TÉCNICO LISBOA

Aprendizagem 2021/22

Homework IV – Group 17

I. Pen-and-paper

1) I.

1) $\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$; $\Sigma_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$; $\pi_1 = P(C_1=1) = 0{,}7$, $\pi_2 = P(C_2=1) = 0{,}3$

$K = 2$

$\mu_1 = x_1 = \begin{pmatrix} 2 \\ 4 \end{pmatrix}$ $\quad \mu_2 = x_2 = \begin{pmatrix} -1 \\ -4 \end{pmatrix}$

• E - step

$C_1 = 1$:

$$p(x \mid C_1=1) = \mathcal{N}(x \mid \mu_1, \Sigma_1) = \frac{1}{(2\pi)^{\frac{3}{2}}} \cdot |\Sigma_1|^{-\frac{1}{2}} \cdot \exp\left(-\frac{1}{2}(x - \mu_1)\Sigma^{-1}(x-\mu_1)^T\right)$$

$p(x_1|C_1=1) = 0{,}1592$ $\quad P(x_1, C_1) = P(x_1|C_1=1) \cdot \pi_1 = 0{,}1114$

$p(x_2|C_1=1) = 2{,}2391 \times 10^{-17}$ $\quad P(x_2, C_1) = P(x_2|C_1=1) \cdot \pi_1 = 1{,}5674 \times 10^{-17}$

$p(x_3|C_1=1) = 0{,}0002383$ $\quad P(x_3, C_1) = P(x_3|C_1=1) \cdot \pi_1 = 0{,}0001675$

$p(x_4|C_1=1) = 7{,}2256 \times 10^{-6}$ $\quad P(x_4, C_1) = P(x_4|C_1=1) \cdot \pi_1 = 5{,}0579 \times 10^{-6}$

C_2: $\quad P(x_1|C_2=1) = 9{,}4388 \times 10^{-10}$ $\quad P(x_1, C_2) = P(x_1|C_2=1) \cdot \pi_2 = 2{,}8316 \times 10^{-10}$

$P(x_2|C_2=1) = 0{,}07958$ $\quad P(x_2, C_2) = P(x_2|C_2=1) \cdot \pi_2 = 0{,}02387$

$P(x_3|C_2=1) = 9{,}8206 \times 10^{-6}$ $\quad P(x_3, C_2) = P(x_3|C_1=1) \cdot \pi_2 = 2{,}9462 \times 10^{-6}$

$P(x_4|C_2=1) = 2{,}8137 \times 10^{-6}$ $\quad P(x_4, C_1) = P(x_4|C_2=1) \cdot \pi_2 = 8{,}4410 \times 10^{-7}$

$P(x_1) = P(x_1, C_1=1) + P(x_1, C_2=1) = 0{,}1114$

$P(x_2) = 0{,}02387$

$P(x_3) = 1{,}7045 \times 10^{-4}$

$P(x_4) = 5{,}902 \times 10^{-6}$

Posteriors:

$P(C_1=1|x_1) = \dfrac{P(C_1=1, x_1)}{P(x_1)} = 1$

$P(C_2=1|x_1) = 0$

$P(C_1=1|x_2) = 0$

$P(C_2=1|x_2) = 1$

$P(C_1=1|x_3) = 0{,}9827$

$P(C_2=1|x_3) = 0{,}0173$

$P(C_1=1|x_4) = 0{,}8570$

$P(C_2=1|x_4) = 0{,}1430$

$\Sigma_1 = \begin{pmatrix} 4{,}1328 & -1{,}1634 \\ -1{,}1634 & 2{,}6056 \end{pmatrix}$

$\Sigma_2 = \begin{pmatrix} 2{,}7014 & 2{,}1060 \\ 2{,}1060 & 2{,}1694 \end{pmatrix}$

• Maximização

$N_1 = \sum_{n=1}^{4} P(C_1=1|x_n) = 2{,}8397$

$N_2 = \sum_{n=1}^{4} P(C_2=1|x_n) = 1{,}1603$

$\pi_1 = P(C_1=1) = \dfrac{N_1}{N} = \dfrac{2{,}8397}{4} = 0{,}7099$

$\pi_2 = P(C_2=1) = \dfrac{N_2}{N} = \dfrac{1{,}1603}{4} = 0{,}2901$

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} P(C_k=1|x_n) \cdot x_n$$

$$\mu_1 = \frac{1}{2{,}8397}\left(1\begin{pmatrix} 2 \\ 4 \end{pmatrix} + 0\begin{pmatrix} -1 \\ -4 \end{pmatrix} + 0{,}9827\begin{pmatrix} -1 \\ 2 \end{pmatrix} + 0{,}8570\begin{pmatrix} 4 \\ 0 \end{pmatrix}\right) = \begin{pmatrix} 1.5654 \\ 2.1007 \end{pmatrix}$$

$$\mu_2 = \frac{1}{1{,}1603}\left(0\begin{pmatrix} 2 \\ 4 \end{pmatrix} + 1\begin{pmatrix} -1 \\ -4 \end{pmatrix} + 0{,}0173\begin{pmatrix} -1 \\ 2 \end{pmatrix} + 0{,}1430\begin{pmatrix} 4 \\ 0 \end{pmatrix}\right) = \begin{pmatrix} -0{,}3838 \\ -3{,}4176 \end{pmatrix}$$

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} P(C_k=1|x_n) \cdot \underbrace{(x_n - \mu_k) \cdot (x_n - \mu_k)^T}$$

$(x_1-\mu_1)(x_1-\mu_1)^T = \begin{pmatrix} 0{,}1889 & 0{,}8254 \\ 0{,}8254 & 3{,}6013 \end{pmatrix}$

$(x_2-\mu_1)(x_2-\mu_1)^T = \begin{pmatrix} 6{,}5813 & 15{,}6507 \\ 15{,}6507 & 37{,}2185 \end{pmatrix}$

$(x_3-\mu_1)(x_3-\mu_1)^T = \begin{pmatrix} 6{,}5813 & 0{,}2583 \\ 0{,}2583 & 0{,}0101 \end{pmatrix}$

$(x_4-\mu_1)(x_4-\mu_1)^T = \begin{pmatrix} 5{,}9273 & -5{,}1144 \\ -5{,}1144 & 4{,}4129 \end{pmatrix}$

$(x_1-\mu_2)(x_1-\mu_2)^T = \begin{pmatrix} 5{,}6825 & 17{,}6821 \\ 17{,}6821 & 55{,}0208 \end{pmatrix}$

$(x_2-\mu_2)(x_2-\mu_2)^T = \begin{pmatrix} 0{,}3797 & 0{,}3589 \\ 0{,}3589 & 0{,}3392 \end{pmatrix}$

$(x_3-\mu_2)(x_3-\mu_2)^T = \begin{pmatrix} 0{,}3797 & -3{,}3383 \\ -3{,}3383 & 29{,}3504 \end{pmatrix}$

$(x_4-\mu_2)(x_4-\mu_2)^T = \begin{pmatrix} 19{,}2177 & 14{,}9821 \\ 14{,}9821 & 11{,}6800 \end{pmatrix}$

$x_1 - \mu_1 = \begin{pmatrix} 2 \\ 4 \end{pmatrix} - \begin{pmatrix} 1{,}5654 \\ 2{,}1007 \end{pmatrix} = \begin{pmatrix} 0{,}4346 \\ 1{,}8993 \end{pmatrix}$

$x_2 - \mu_1 = \begin{pmatrix} -1 \\ -4 \end{pmatrix} - \begin{pmatrix} 1{,}5654 \\ 2{,}1007 \end{pmatrix} = \begin{pmatrix} -2{,}5654 \\ -6{,}1007 \end{pmatrix}$

$x_3 - \mu_1 = \begin{pmatrix} -1 \\ 2 \end{pmatrix} - \begin{pmatrix} 1{,}5654 \\ 2{,}1007 \end{pmatrix} = \begin{pmatrix} -2{,}5654 \\ -0{,}1007 \end{pmatrix}$

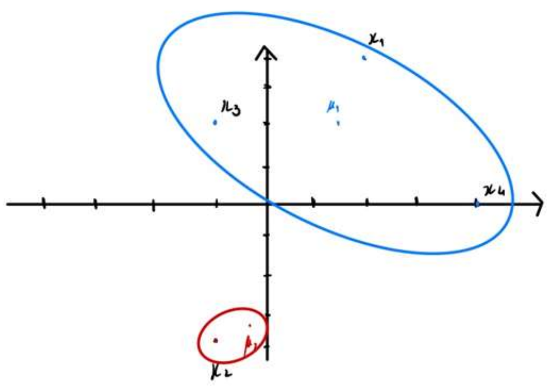
$x_4 - \mu_1 = \begin{pmatrix} 4 \\ 0 \end{pmatrix} - \begin{pmatrix} 1{,}5654 \\ 2{,}1007 \end{pmatrix} = \begin{pmatrix} 2{,}4346 \\ -2{,}1007 \end{pmatrix}$

$x_1 - \mu_2 = \begin{pmatrix} 2 \\ 4 \end{pmatrix} - \begin{pmatrix} -0{,}3838 \\ -3{,}4176 \end{pmatrix} = \begin{pmatrix} 2{,}3838 \\ 7{,}4176 \end{pmatrix}$

$x_2 - \mu_2 = \begin{pmatrix} -1 \\ -4 \end{pmatrix} - \begin{pmatrix} -0{,}3838 \\ -3{,}4176 \end{pmatrix} = \begin{pmatrix} -0{,}6162 \\ -0{,}6824 \end{pmatrix}$

$x_3 - \mu_2 = \begin{pmatrix} -1 \\ 2 \end{pmatrix} - \begin{pmatrix} -0{,}3838 \\ -3{,}4176 \end{pmatrix} = \begin{pmatrix} -0{,}6162 \\ 5{,}4176 \end{pmatrix}$

$x_4 - \mu_2 = \begin{pmatrix} 4 \\ 0 \end{pmatrix} - \begin{pmatrix} -0{,}3838 \\ -3{,}4176 \end{pmatrix} = \begin{pmatrix} 4{,}3838 \\ 3{,}4176 \end{pmatrix}$



Aprendizagem 2021/22

Homework IV – Group 17

2) 2) $s = 1 - \frac{a}{b}$

$a_1 = \frac{\|x_1 - x_3\|_2 + \|x_1 - x_4\|_2}{2} = \frac{3,6056 + 4,4721}{2} = 4,0389$

$a_3 = \frac{\|x_3 - x_1\|_2 + \|x_3 - x_4\|_2}{2} = \frac{3,6056 + 5,3852}{2} = 4,4954$

$a_4 = \frac{\|x_4 - x_1\|_2 + \|x_4 - x_3\|_2}{2} = \frac{4,4721 + 5,3852}{2} = 4,9287$

$s_1 = 1 - \frac{a_1}{b_1} = 0,5273$

$s_2 = 0$, por definição pois $|C_2| = 1$.

$s_3 = 1 - \frac{a_3}{b_3} = 0,2507$

$s_4 = 1 - \frac{a_4}{b_4} = 0,2303$

$b_1 = \|x_1 - x_2\|_2 = 8,5440$

$b_3 = \|x_3 - x_2\|_2 = 6$

$b_4 = \|x_4 - x_2\|_2 = 6,4031$

$s_{C_2} = 0$; $s_{C_1} = \frac{s_1 + s_3 + s_4}{3} = \frac{0,5273 + 0,2507 + 0,2303}{3} = 0,3361$

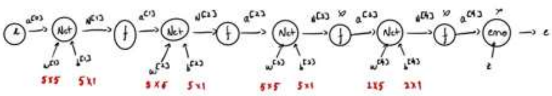
$s = \frac{0 + 0,3361}{2} = 0,1681$

Verifica-se que a silhueta obtida é bastante baixa (muito inferior a 1), logo seriam necessários mais clusters para melhor classificação dos pontos.

3)

a) 3)

a) MLP, inner layers: $(5,5,5)$



VC-Dimension = $3 \times 5^2 + 3 \times 5 + 2 \times 5 + 2 \times 1 = 102$

$3 \times n^2 + 3 \times n + 2 \times n + 2$

Decision tree, 3 bins:

VC-Dimension = $3^d = 3^5 = 243$

3^n

Bayes Classifier, multivariate Gauss likelihood

$\left.\begin{array}{l} \mu \rightarrow 5 \text{ parâmetros} \\ \Sigma \rightarrow \frac{5^2 - 5}{2} + 5 = 15 \end{array}\right\}$ Por classe

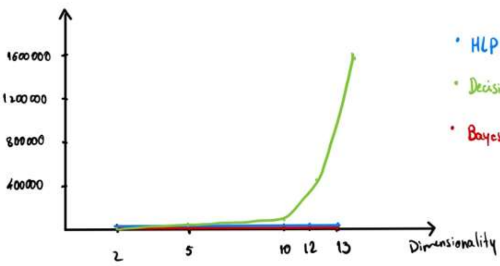
$P(C) \rightarrow 1$ parâmetro (apenas 2 classes)

$\left(\frac{n^2 + n}{2} + n\right) \times 2 + 1$

VC-Dimension $= (15 + 5) \times 2 + 1 = 41$

b)

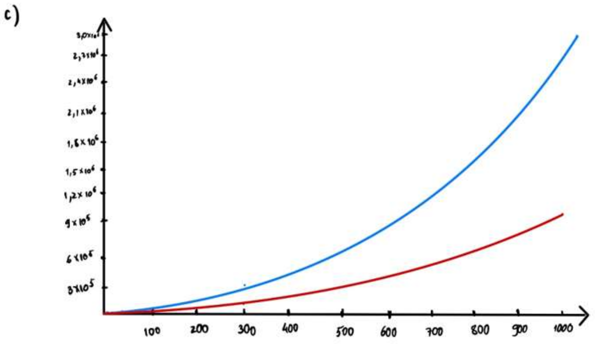
b)



- MLP $\{(2,24), (5,102), (10,352), (12,494), (13,574)\}$
- Decision Tree $\{(2,9), (5,243), (10,59049), (12,531441), (13,1594323)\}$
- Bayesian classifier $\{(2,11), (5,41), (10,131), (12,181), (13,209)\}$

É possível concluir que com o aumento da dimensionalidade da amostra, a VC-Dimension da decision tree tem um crescimento muito mais acentuado que tanto o MLP e o Bayes Classifier. $(O(3^d) >> O(d^2))$

c)



• MLP {(2,24), (5,102), (10,352), (30,2852), (100,30502), (300,271502), (1000,3005002)}

• Bayesian classifier {(2,11), (5,41), (10,131), (30,991), (100,10301), (300,90901), (1000,1003001)}

Apesar de ambos os modelos terem VC-Dimension de crescimento quadrático, o MLP tem maior crescimento e consequentemente maior VC-Dimension.

II. Programming and critical analysis

4)

a) k = 2: ECR = 13.5

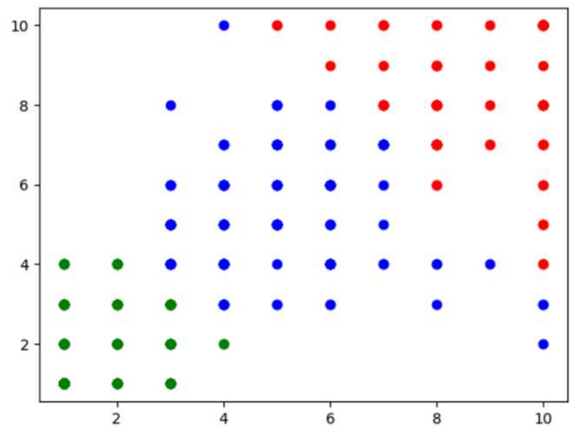
k = 3: ECR = 6.67

b) k = 2: silhouette = 0.5968

k = 3: silhouette = 0.5245

A silhueta é um critério interno de medição que avalia a separação e a coesão da solução de clustering, enquanto o ECR é um critério externo que avalia o quão bem os clusters se ajustam à classificação real das observações. Assim, comparando as silhuetas verifica-se que com k = 2 obtêm-se clusters melhor separados e mais coesos, mas, tendo em conta o ECR, verificamos que estes agrupamentos não são homogéneos, sendo, de acordo com esta métrica, a solução com k = 3 melhor.

5)



6) Ao observarmos o gráfico obtido, verificamos que as 2 features com maior mutual information produzem uma boa solução de clustering, uma vez que se observa uma boa coesão dos pontos pertencentes ao mesmo cluster e uma boa separação entre os clusters.

III. APPENDIX

```python
import numpy as np
import scipy.io.arff
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
from sklearn.feature_selection import SelectKBest, mutual_info_classif

arff = scipy.io.arff

breast_data, breast_meta = arff.loadarff("breast.w.new.arff")

breast_targets = np.array(list(el["Class"] for el in breast_data))
breast_X = np.array(np.array(list(list(el[i] for i in range(9)) for el in breast_data)))

clusters2 = KMeans(n_clusters=2)
clusters3 = KMeans(n_clusters=3)

clusters2.fit(breast_X)
clusters3.fit(breast_X)

#4a

benign2 = [0, 0]
benign3 = [0, 0, 0]
malign2 = [0, 0]
malign3 = [0, 0, 0]

for i in range(len(clusters2.labels_)):
    if breast_targets[i] == b'benign':
        benign2[clusters2.labels_[i]] += 1
    else:
        malign2[clusters2.labels_[i]] += 1

for i in range(len(clusters3.labels_)):
    if breast_targets[i] == b'benign':
        benign3[clusters3.labels_[i]] += 1
    else:
        malign3[clusters3.labels_[i]] += 1

total_cluster_2_0 = benign2[0] + malign2[0]
total_cluster_2_1 = benign2[1] + malign2[1]

total_cluster_3_0 = benign3[0] + malign3[0]
total_cluster_3_1 = benign3[1] + malign3[1]
total_cluster_3_2 = benign3[2] + malign3[2]
```

```
ecr2 = 0.5 * ((total_cluster_2_0 - max(benign2[0], malign2[0])) + (total_cluster_2_1 -
max(benign2[1], malign2[1])))
ecr3 = (1/3) * ((total_cluster_3_0 - max(benign3[0], malign3[0])) + (total_cluster_3_1 -
max(benign3[1], malign3[1])) + (total_cluster_3_2 - max(benign3[2], malign3[2])))

print(ecr2, ecr3)

#4b

silhouette2 = silhouette_score(breast_X, clusters2.labels_)
silhouette3 = silhouette_score(breast_X, clusters3.labels_)

print(silhouette2, silhouette3)

#5

kbest = SelectKBest(mutual_info_classif, k=2).fit_transform(breast_X, breast_targets)

clusters = KMeans(n_clusters=3)
clusters.fit(kbest)

clusters_div = [[], [], []]

for i in range(len(clusters.labels_)):
    clusters_div[clusters.labels_[i]].append(kbest[i])

for i in range(len(clusters_div)):
    clusters_div[i] = np.array(clusters_div[i])

plt.scatter(clusters_div[0][:,0], clusters_div[0][:,1], color="red")
plt.scatter(clusters_div[1][:,0], clusters_div[1][:,1], color="green")
plt.scatter(clusters_div[2][:,0], clusters_div[2][:,1], color="blue")

plt.savefig("graphic.png", format="png")
```

END