TÉCNICO LISBOA

Aprendizagem 2021/22

Homework I – Group 057

I. Pen-and-paper

	y1	y3	y4
média	0,13	0,15	0,15
média | 0	0,25	0,2	0,25
média | 1	0,05	0,116666667	0,083333
desv. P.	0,275076		
desv. P. | 0	0,238048		
desv. P. | 1	0,288097		
var | 0		0,18	0,25
var | 1		0,109666667	0,213667
cov | 0		0,18	
cov | 1		0,122333333	
cov		0,131666667	
var		0,122777778	0,209444

1) $P(N) = \frac{4}{10}$ $\qquad P(P) = \frac{6}{10}$

$$y_1 \sim N(0{,}13\,;\,0{,}275076^2)$$
$$y_1 | N \sim N(0{,}25\,;\,0{,}238048^2)$$
$$y_1 | P \sim N(0{,}05\,;\,0{,}288097^2)$$

$$P(y_2 = A | N) = \frac{2}{4} \qquad P(y_2 = A | P) = 1/6$$
$$P(y_2 = B | N) = \frac{1}{4} \qquad P(y_2 = B | P) = 2/6$$
$$P(y_2 = C | N) = \frac{1}{4} \qquad P(y_2 = C | P) = 3/6$$

$$(y_3, y_4): \quad \mu = \begin{bmatrix} 0{,}15 \\ 0{,}15 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 0{,}122778 & 0{,}1316667 \\ 0{,}1316667 & 0{,}209444 \end{bmatrix}$$

$$(y_3, y_4) | N: \quad \mu = \begin{bmatrix} 0{,}2 \\ 0{,}25 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 0{,}18 & 0{,}18 \\ 0{,}18 & 0{,}25 \end{bmatrix} \quad |\Sigma| = 0{,}0126$$

$$(y_3, y_4) | P: \quad \mu = \begin{bmatrix} 0{,}116667 \\ 0{,}083333 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 0{,}109667 & 0{,}122333 \\ 0{,}122333 & 0{,}213667 \end{bmatrix} \quad |\Sigma| = 0{,}008466856$$

$$P(N | x_{new}) = \frac{P(x_{new} | N)\, P(N)}{P(x_{new})} \qquad x_{new} = [x_1, x_2, x_3, x_4]$$

$$P(x_{new} | N) = P(y_1 = x_1, y_2 = x_2, y_3 = x_3, y_4 = x_4 | N) =$$
$$= P(y_1 = x_1 | N)\, P(y_2 = x_2 | N)\, P(y_3 = x_3, y_4 = x_4 | N)$$

$$\boxed{P(N | x_{new}) = \frac{\underbrace{P(y_1 = x_1 | N)\, P(y_2 = x_2 | N)\, P(y_3 = x_3, y_4 = x_4 | N)}}{P(x_{new})} \times P(N)}$$

$$P(y_1 = x_1 | N) = N(x_1 | \mu = 0{,}25, \sigma = 0{,}238048)$$
$$= \frac{1}{\sqrt{2\pi}} \frac{1}{0{,}238048} \exp\left(-\frac{1}{2 \times 0{,}238048^2} \cdot (x_1 - 0{,}25)^2 \right)$$

$$P(y_3 = x_3, y_4 = x_4 | N) = N\left(\begin{bmatrix} x_3 \\ x_4 \end{bmatrix} \middle| \mu = \begin{bmatrix} 0{,}2 \\ 0{,}25 \end{bmatrix}, \Sigma = \begin{bmatrix} 0{,}18 & 0{,}18 \\ 0{,}18 & 0{,}25 \end{bmatrix} \right) =$$

$$= \frac{1}{2\pi} \frac{1}{\sqrt{0{,}0126}} \exp\left(-\frac{1}{2} \left(\begin{bmatrix} x_3 \\ x_4 \end{bmatrix} - \begin{bmatrix} 0{,}2 \\ 0{,}25 \end{bmatrix} \right)^T \times \begin{bmatrix} 0{,}18 & 0{,}18 \\ 0{,}18 & 0{,}25 \end{bmatrix}^{-1} \left(\begin{bmatrix} x_3 \\ x_4 \end{bmatrix} - \begin{bmatrix} 0{,}2 \\ 0{,}25 \end{bmatrix} \right) \right)$$

$$\boxed{P(P \mid x_{new}) = \frac{P(y_1 = x_1 \mid P)\, P(y_2 = x_2 \mid P)\, P(y_3 = x_3, y_4 = x_4 \mid P)}{P(x_{new})} \times P(P)}$$

$$P(y_1 = x_1 \mid N) = N(x_1 \mid \mu = 0{,}05, \sigma = 0{,}288097)$$

$$= \frac{1}{\sqrt{2\pi}} \frac{1}{0{,}288097} \exp\left(-\frac{1}{2 \times 0{,}288097^2} \cdot (x_1 - 0{,}05)^2\right)$$

$$P(y_3 = x_3, y_4 = x_4 \mid N) = N\left(\begin{bmatrix} x_3 \\ x_4 \end{bmatrix} \middle| \mu = \begin{bmatrix} 0{,}116667 \\ 0{,}083333 \end{bmatrix}, \Sigma = \begin{bmatrix} 0{,}109667 & 0{,}122333 \\ 0{,}122333 & 0{,}213667 \end{bmatrix}\right)$$

$$= \frac{1}{2\pi} \frac{1}{\sqrt{0{,}0126}} \exp\left(-\frac{1}{2}\left(\begin{bmatrix} x_3 \\ x_4 \end{bmatrix} - \begin{bmatrix} 0{,}116667 \\ 0{,}083333 \end{bmatrix}\right)^T \begin{bmatrix} 0{,}109667 & 0{,}122333 \\ 0{,}122333 & 0{,}213667 \end{bmatrix}^{-1} \left(\begin{bmatrix} x_3 \\ x_4 \end{bmatrix} - \begin{bmatrix} 0{,}116667 \\ 0{,}083333 \end{bmatrix}\right)\right)$$

2) $[0{,}6 ; A ; 0{,}2 ; 0{,}4] = x_{new} = x_1$

- $P(x_{new} \mid N) = P(y_1 = 0{,}6 \mid N) \underbrace{P(y_2 = A \mid N)}_{2/4} P(y_3 = 0{,}2 ; y_4 = 0{,}4 \mid N)$

$$= \frac{1}{\sqrt{2\pi}} \frac{1}{0{,}238048} \exp\left(-\frac{1}{2(0{,}238048)^2} \cdot (0{,}6 - 0{,}25)^2\right) \times \frac{2}{4} \times \frac{1}{2\pi} \frac{1}{\sqrt{0{,}0126}} \exp\left(-\frac{1}{2} \cdot \left(\begin{bmatrix} 0{,}2 \\ 0{,}4 \end{bmatrix} - \begin{bmatrix} 0{,}2 \\ 0{,}25 \end{bmatrix}\right)^T \frac{1}{0{,}0126} \begin{bmatrix} 0{,}25 & -0{,}18 \\ -0{,}18 & 0{,}18 \end{bmatrix} \left(\begin{bmatrix} 0{,}2 \\ 0{,}4 \end{bmatrix} - \begin{bmatrix} 0{,}2 \\ 0{,}25 \end{bmatrix}\right)\right) =$$

$$= 0{,}403117 \exp\left(-\frac{1}{2} \begin{bmatrix} 0 \\ 0{,}15 \end{bmatrix}^T \frac{1}{0{,}0126} \begin{bmatrix} 0{,}25 & -0{,}18 \\ -0{,}18 & 0{,}18 \end{bmatrix} \begin{bmatrix} 0 \\ 0{,}15 \end{bmatrix}\right) =$$

$$= 0{,}403117 \exp\left(\frac{-1}{2 \times 0{,}0126} \begin{bmatrix} 0 & 0{,}15 \end{bmatrix} \begin{bmatrix} 0{,}25 & -0{,}18 \\ -0{,}18 & 0{,}18 \end{bmatrix} \begin{bmatrix} 0 \\ 0{,}15 \end{bmatrix}\right) =$$

$$= 0{,}403117 \exp\left(\frac{-1}{0{,}0252} \begin{bmatrix} -0{,}027 & 0{,}027 \end{bmatrix} \begin{bmatrix} 0 \\ 0{,}15 \end{bmatrix}\right) =$$

$$= 0{,}403117 \exp\left(\frac{-1}{0{,}0252} \times 0{,}00405\right) =$$

$$= 0{,}343268$$

$$P(N \mid x_{new}) = \frac{0{,}343268 \times P(N)}{P(x_{new})}$$

- $P(x_{new} \mid P) = P(y_1 = 0{,}6 \mid P) \underbrace{P(y_2 = A \mid P)}_{1/6} P(y_3 = 0{,}2 ; y_4 = 0{,}4 \mid P)$

$$= \frac{1}{\sqrt{2\pi}} \frac{1}{0{,}288097} \exp\left(\frac{-1}{2(0{,}288097)^2} (0{,}6 - 0{,}05)^2\right) \times \frac{1}{6} \times \frac{1}{2\pi} \frac{1}{\sqrt{0{,}008466856}} \exp\left(-\frac{1}{2}\left(\begin{bmatrix} 0{,}2 \\ 0{,}4 \end{bmatrix} - \begin{bmatrix} 0{,}116667 \\ 0{,}083333 \end{bmatrix}\right)^T \frac{1}{0{,}008466856} \begin{bmatrix} 0{,}213667 & -0{,}122333 \\ -0{,}122333 & 0{,}109667 \end{bmatrix} \left(\begin{bmatrix} 0{,}2 \\ 0{,}4 \end{bmatrix} - \begin{bmatrix} 0{,}116667 \\ 0{,}083333 \end{bmatrix}\right)\right) =$$

$$= 0{,}064531 \exp\left(-\frac{1}{2} \begin{bmatrix} 0{,}083333 & 0{,}316667 \end{bmatrix} \frac{1}{0{,}008466856} \begin{bmatrix} 0{,}213667 & -0{,}122333 \\ -0{,}122333 & 0{,}109667 \end{bmatrix} \begin{bmatrix} 0{,}083333 \\ 0{,}316667 \end{bmatrix}\right) =$$

$$= 0{,}064531 \exp\left(\frac{-1}{0{,}016934} \begin{bmatrix} -0{,}020933 & 0{,}024534 \end{bmatrix} \begin{bmatrix} 0{,}083333 \\ 0{,}316667 \end{bmatrix}\right) =$$

$$= 0{,}045212$$

$$P(P \mid x_{new}) = \frac{0{,}045212 \times P(P)}{P(x_{new})}$$

- Comparando $P(N \mid x_{new})$ com $P(P \mid x_{new})$ . . .

$$= \frac{0{,}137307}{P(x_{new})} \qquad = \frac{0{,}027127}{P(x_{new})}$$

visto que têm o mesmo denominador, $P(N \mid x_{new}) > P(P \mid x_{new})$ logo o classificador atribui a x_{new} a classe N

Efetuando os mesmos cálculos acima para os restantes x e verificando os resultados com o programa Python presente no Apêndice, conseguimos completar a seguinte tabela: (0=N e 1=P)

xi	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
True Class	0	0	0	0	1	1	1	1	1	1
P(xi|N)P(N)	0,137	0,063	0,232	0,070	0,193	0,019	0,008	0,178	0,060	0,030
P(xi|P)P(P)	0,027	0,261	0,074	0,083	0,229	0,243	0,121	0,203	0,026	0,321
Predicted Class	0	1	0	1	1	1	1	1	0	1

Predicted

True		N	P
	N	2	2
	P	1	5

3) $$F_1 = 2\frac{1}{\frac{1}{recall}+\frac{1}{precisão}}$$

Recall = 5/(5+1) = 0,8333333333
Precisão = 5/(5+2) = 0,7142857143
F1 = 2/(1/0,8333333333+1/0,7142857143) = 0,7692307692

4)

P(xi) = P(xi|N)P(N) + P(xi|P)P(P)
P(N|xi) = P(xi|N)P(N) / P(xi)

xi	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
True Class	0	0	0	0	1	1	1	1	1	1
P(N|xi)	0,835	0,194	0,758	0,458	0,457	0,072	0,062	0,467	0,698	0,085
Threshold = 70%	N	P	N	P	P	P	P	P	P	P

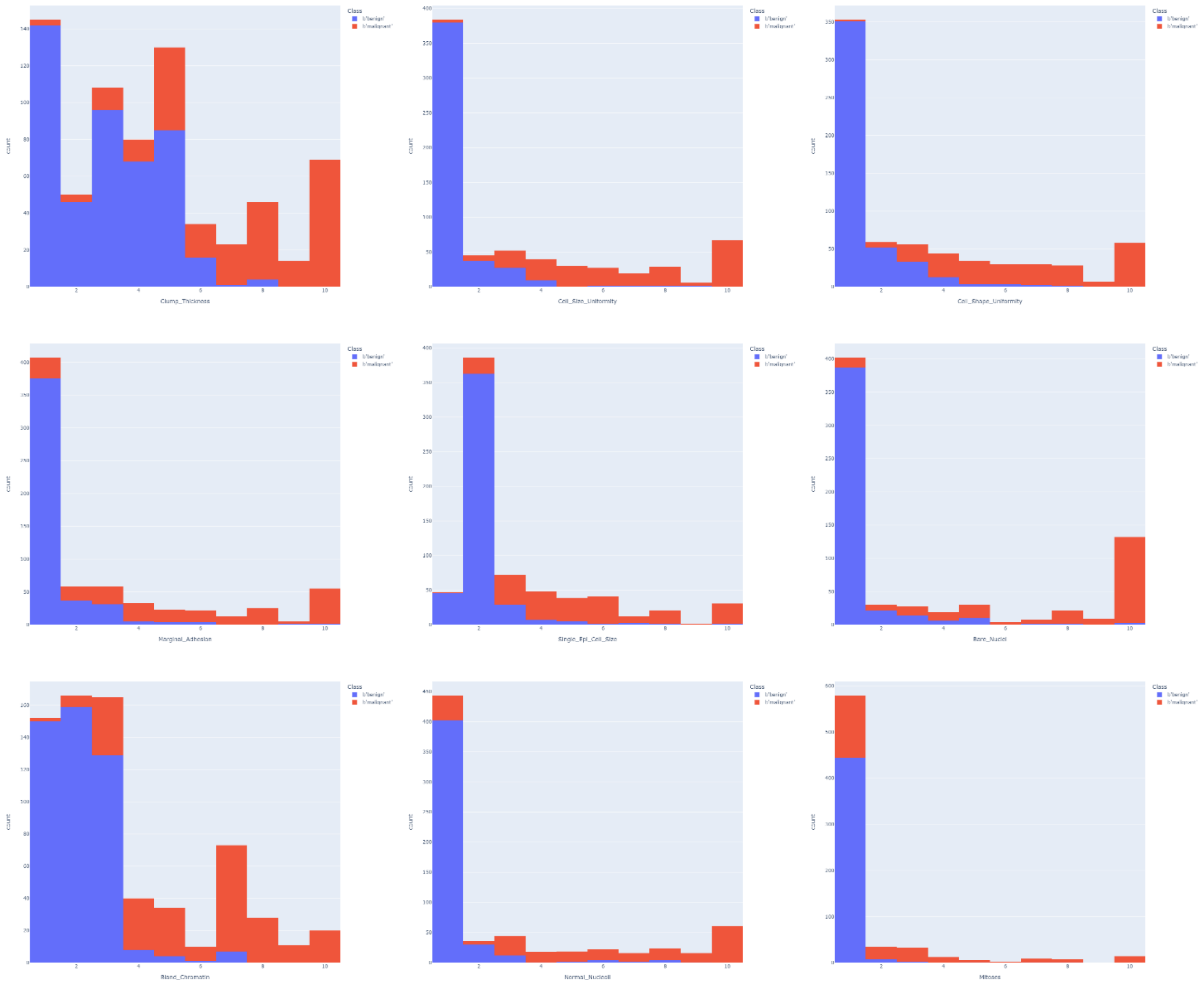
Após testarmos com vários thresholds diferentes, concluímos que a melhor accuracy possível é de 8/10 = 80 %, utilizando como thresholds valores entre os 70 e 75%.

Aprendizagem 2021/22

Homework I – Group 057

II. Programming and critical analysis

5) azul – benign, vermelho - malignant



6) k=3 → Accuracy = 96,7882 %
k=5 → Accuracy = 97,2272 %
k=7 → Accuracy = 97,0801 %
Uma vez que k=5 tem a maior accuracy, 5 é o valor de k que está menos suscetível a overfitting.

7) kNN com k=3 e 10-fold cross validation com seed=57 → Accuracy = 96,7882 %
Naïve Bayes, também com 10-fold cross validation e seed=57 → Accuracy = 95.9064 %
Assim, concluimos que a hipótese de que kNN é estatisticamente superior a Naïve Bayes encontra-se correta para este data set específico.

8) Uma das desvantagens do Naïve Bayes relativamente ao kNN é que Naïve Bayes assume que todas as variáveis são independentes entre si, o que na prática não se verifica completamente, afetando negativamente a accuracy. Além disso, o Naïve Bayes, por ser um classificador mais apreensivo, acaba por ser mais rápido mas menos preciso. O kNN devido à sua natureza inerente para otimizar localmente acaba por ter resultados mais precisos.

III. APPENDIX

```
# Código parte I

import math
import numpy as np

y1_u = 0.13
y1_o = 0.275076
y1_uN = 0.25
y1_oN = 0.238048
y1_uP = 0.05
y1_oP = 0.288097
y34_u = np.array([[0.15],
                  [0.15]])
y34_o = np.array([[0.122778, 0.1316667],
                 [0.1316667, 0.209444]])
y34_uN = np.array([[0.2],
                  [0.25]])
y34_oN = np.array([[0.18, 0.18],
                  [0.18, 0.25]])
y34_uP = np.array([[0.116667],
                  [0.083333]])
y34_oP = np.array([[0.109667, 0.122333],
                  [0.122333, 0.213667]])

def dividendo_pNx (y1, y2, y3, y4):
    det = np.linalg.det(y34_oN)
    res = 1/(y1_oN) * 1/math.sqrt(2*math.pi)
    sub = np.subtract(np.array([[y3],[y4]]), y34_uN)

    if y2 == 'A':
        p = 0.5
    else:
        p = 0.25

    res = res * p * (1/(2*math.pi)) * (1/math.sqrt(det))
    res = res * math.exp( (-1/(2*y1_oN**2)) * (y1-y1_uN)**2)
    matrixes = np.matmul(np.transpose(sub), np.linalg.inv(y34_oN))
    matrixes = np.matmul(matrixes, sub)
    res = res * math.exp(-matrixes/2)
    return res*0.4

def dividendo_pPx (y1, y2, y3, y4):
    det = np.linalg.det(y34_oP)
    res = 1/(y1_oP) * 1/math.sqrt(2*math.pi)
    sub = np.subtract(np.array([[y3],[y4]]), y34_uP)

    if y2 == 'A':
        p = 1.0/6
    elif y2 == 'B':
        p = 2.0/6
    else:
        p = 3.0/6

    res = res * p * (1/(2*math.pi)) * (1/math.sqrt(det))
    res = res * math.exp( (-1/(2*y1_oP**2)) * (y1-y1_uP)**2)
    matrixes = np.matmul(np.transpose(sub), np.linalg.inv(y34_oP))
    print (matrixes)
    matrixes = np.matmul(matrixes, sub)
    res = res * math.exp(-matrixes/2)
    return res*0.6
```

```
# Código parte II

import plotly.express as px
import pandas as pd
from scipy.io import arff

from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score
from sklearn.neighbors import KNeighborsClassifier
from sklearn import metrics
from sklearn.naive_bayes import GaussianNB

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')

X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=57)

knn = KNeighborsClassifier(n_neighbors=3) # mudar 3 para o k desejado
scores = cross_val_score(knn, X, y, cv=10, scoring='accuracy')
print(scores)
print("Accuracy kNN:", scores.mean())

gnb = GaussianNB()
gnb.fit(X_train, y_train)
y_pred = gnb.predict(X_test)
print("Accuracy Naive Bayes:", metrics.accuracy_score(y_test, y_pred))

# Histogramas
fig = px.histogram(df, x="Clump_Thickness", color="Class")
fig.show()
fig = px.histogram(df, x="Cell_Size_Uniformity", color="Class")
fig.show()
fig = px.histogram(df, x="Cell_Shape_Uniformity", color="Class")
fig.show()
fig = px.histogram(df, x="Marginal_Adhesion", color="Class")
fig.show()
fig = px.histogram(df, x="Single_Epi_Cell_Size", color="Class")
fig.show()
fig = px.histogram(df, x="Bare_Nuclei", color="Class")
fig.show()
fig = px.histogram(df, x="Bland_Chromatin", color="Class")
fig.show()
fig = px.histogram(df, x="Normal_Nucleoli", color="Class")
fig.show()
fig = px.histogram(df, x="Mitoses", color="Class")
fig.show()
```

* breast.w_modified.arff é o dataset fornecido, mas sem as 16 observações com valores em falta

END