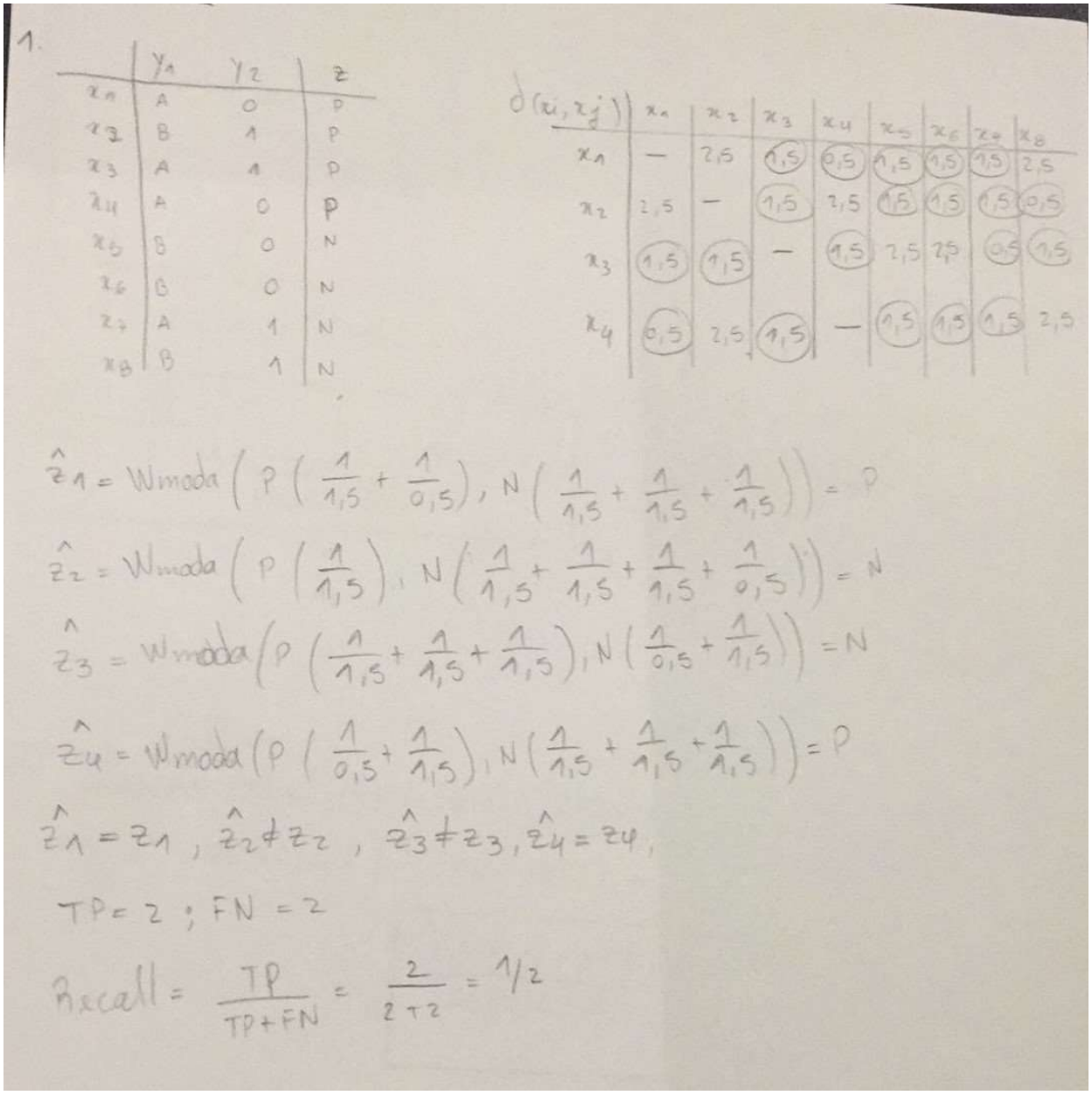
Aprendizagem 2022/23

Homework I – Group 084

I. Pen-and-paper

1)



1.

	Y_1	Y_2	Z
x_1	A	0	P
x_2	B	1	P
x_3	A	1	P
x_4	A	0	P
x_5	B	0	N
x_6	B	0	N
x_7	A	1	N
x_8	B	1	N

$d(x_i, x_j)$	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
x_1	—	2,5	1,5	0,5	1,5	1,5	1,5	2,5
x_2	2,5	—	1,5	2,5	1,5	1,5	1,5	0,5
x_3	1,5	1,5	—	1,5	2,5	2,5	0,5	1,5
x_4	0,5	2,5	1,5	—	1,5	1,5	1,5	2,5

$\hat{z}_1 = \text{Wmoda}\left(P\left(\frac{1}{1,5} + \frac{1}{0,5}\right), N\left(\frac{1}{1,5} + \frac{1}{1,5} + \frac{1}{1,5}\right)\right) = P$

$\hat{z}_2 = \text{Wmoda}\left(P\left(\frac{1}{1,5}\right), N\left(\frac{1}{1,5} + \frac{1}{1,5} + \frac{1}{1,5} + \frac{1}{0,5}\right)\right) = N$

$\hat{z}_3 = \text{Wmoda}\left(P\left(\frac{1}{1,5} + \frac{1}{1,5} + \frac{1}{1,5}\right), N\left(\frac{1}{0,5} + \frac{1}{1,5}\right)\right) = N$

$\hat{z}_4 = \text{Wmoda}\left(P\left(\frac{1}{0,5} + \frac{1}{1,5}\right), N\left(\frac{1}{1,5} + \frac{1}{1,5} + \frac{1}{1,5}\right)\right) = P$

$\hat{z}_1 = z_1$, $\hat{z}_2 \neq z_2$, $\hat{z}_3 \neq z_3$, $\hat{z}_4 = z_4$,

$TP = 2$; $FN = 2$

$\text{Recall} = \frac{TP}{TP + FN} = \frac{2}{2+2} = 1/2$

2)

2

	Y_1	Y_2	Y_3	Z
x_1	A	0	1,2	P
x_2	B	1	0,8	P
x_3	A	1	0,5	P
x_4	A	0	0,9	P
x_5	B	0	0,8	P
x_6	B	0	1	N
x_7	B	0	0,9	N
x_8	A	1	1,2	N
x_9	B	1	0,8	N

$P(Y_3 | Z=P)$:

$\mu = \frac{1,2+0,8+0,5+0,9+0,8}{5} = 0,84$

$\sigma^2 = \frac{1}{5-1} \sum_{i=1}^{5} (y_{3i} - \mu)^2 =$

$= \frac{1}{4}\left[(1,2-0,84)^2 + (0,8-0,84)^2 + (0,5-0,84)^2 + (0,9-0,84)^2 + (0,8-0,84)^2\right] =$

$= 0,0635$

$$P(y_3 | z=P) = \frac{1}{\sqrt{2\pi \cdot 0,0635}} \times e^{-\frac{(y_3-0,84)^2}{2 \cdot 0,0635}}$$

$P(Y_3 | Z=N)$:

$\mu = \frac{1+0,9+1,2+0,8}{4} = 0,975$

$\sigma^2 = \frac{1}{4-1} \sum_{i=6}^{9} (y_{3i} - \mu)^2 = \frac{1}{3}\left((1-0,975)^2 + (0,9-0,975)^2 + (1,2-0,975)^2 + (0,8-0,975)^2\right) = 0,0292$

$$P(y_3 | z=N) = \frac{1}{\sqrt{2\pi \cdot 0,0292}} \times e^{-\frac{(y_3-0,975)^2}{2 \cdot 0,0292}}$$

$$P(y_1, y_2 | z=P) = \begin{cases} 2/5, & y_1 = A \wedge y_2 = 0 \\ 1/5, & y_1 = A \wedge y_2 = 1 \\ 1/5, & y_1 = B \wedge y_2 = 0 \\ 1/5, & y_1 = B \wedge y_2 = 1 \\ 0, & \text{c.c.} \end{cases}$$

$$P(y_1, y_2 | z=N) = \begin{cases} 1/4, & y_1 = A \wedge y_2 = 1 \\ 2/4, & y_1 = B \wedge y_2 = 0 \\ 1/4, & y_1 = B \wedge y_2 = 1 \\ 0, & \text{c.c.} \end{cases}$$

$P(z=P) = 5/9$; $P(z=N) = 4/9$

(continuazione)

$$P(z=N|x) = \frac{P(x|z=N) \cdot P(z=N)}{P(x)} = \frac{P(x|z=N) \cdot P(z=N)}{P(x|z=P) \cdot P(z=P) + P(x|z=N) \cdot P(z=N)}$$

$P(x|z=P) = P(y_1, y_2 | z=P) \cdot P(y_3 | z=P)$

$P(x|z=N) = P(y_1, y_2 | z=N) \cdot P(y_3 | z=N)$

$P(z=P|x) = 1 - P(z=N|x)$

3)

3 De acordo com a MAP assumption ao calcularmos $P(\text{Positivo} \mid x)$ queremos apenas maximizar $P(x \mid z=P) \cdot P(z=P)$, dado que o denominador tem apenas como função normalização de valor

$x = \begin{pmatrix} A \\ 1 \\ 0,8 \end{pmatrix}$:

$P(z=P \mid x) = P(x \mid z=P) \cdot P(z=P) =$

$= P(y_1=A, y_2=1 \mid z=P) \cdot P(y_3=0,8 \mid z=P) \cdot P(z=P) =$

$= \frac{1}{5} \cdot \frac{1}{\sqrt{2\pi \cdot 0,0635}} e^{\frac{-(0,8-0,84)^2}{2 \cdot 0,0635}} \cdot 5/9 = 0,174$

$x = \begin{pmatrix} B \\ 1 \\ 1 \end{pmatrix}$:

$P(z=P \mid x) = P(x \mid z=P) \cdot P(z=P) = P(y_1=B, y_2=1 \mid z=P) \cdot P(y_3=1 \mid z=P) \cdot P(z=P)$

$= 1/5 \cdot \frac{1}{\sqrt{2\pi \cdot 0,0635}} e^{\frac{-(1-0,84)^2}{2 \cdot 0,0635}} \cdot 5/9 = 0,144$

$x = \begin{pmatrix} B \\ 0 \\ 0,9 \end{pmatrix}$:

$P(z=P \mid x) = P(x \mid z=P) \cdot P(z=P) \cdot P(y_1=B, y_2=0 \mid z=P) \cdot P(y_3=0,9 \mid z=P) \cdot P(z=P) =$

$= \frac{1}{5} \cdot \frac{1}{\sqrt{2\pi \cdot 0,0635}} e^{\frac{-(0,9-0,84)^2}{2 \cdot 0,0635}} \cdot 5/9 = 0,171$

4)

4.

$$P(z=P|x) = \frac{P(x|z=P)\cdot P(z=P)}{P(x)} =$$

$$= \frac{P(x|z=P)\cdot P(z=P)}{P(x|z=P)\cdot P(z=P) + P(x|z=N)\cdot P(z=N)} =$$

$$= \frac{P(y_3=y_3'|z=P)\cdot P(y_1=y_1', y_2=y_2'|z=P)\cdot P(z=P)}{P(y_3=y_3'|z=P)\cdot P(y_1=y_1', y_2=y_2'|z=P)\cdot P(z=P) + P(y_3=y_3'|z=N)\cdot P(y_1=y_1', y_2=y_2'|z=N)\cdot P(z=N)}$$

$x_1 = \begin{pmatrix} A \\ 1 \\ 0,8 \end{pmatrix}$:

$$\frac{P(y_3=0,8|z=P)\cdot P(y_1=A, y_2=1|z=P)\cdot P(z=P)}{P(y_3=0,8|z=P)\cdot P(y_1=A, y_2=1|z=P)\cdot P(z=P) + P(y_3=0,8|z=N)\cdot P(y_1=A, y_2=1|z=N)\cdot P(z=N)} =$$

$$= \frac{\frac{1}{\sqrt{2\pi}\cdot 0,0635} \times e^{\frac{-(0,8-0,84)^2}{2\cdot 0,0635}} \times \frac{1}{5} \times 5/9}{\frac{1}{\sqrt{2\pi}\cdot 0,0635} \times e^{\frac{-(0,8-0,84)^2}{2\cdot 0,0635}} \times \frac{1}{5} \times \frac{5}{9} + \frac{1}{\sqrt{2\pi}\cdot 0,0292} \times e^{\frac{-(0,8-0,9375)^2}{2\cdot 0,0292}} \times \frac{1}{4} \times \frac{4}{9}} = 0,53$$

$x_2 = \begin{pmatrix} B \\ 1 \\ 1 \end{pmatrix}$:

$$\frac{P(y_3=1|z=P)\cdot P(y_1=B, y_2=1|z=P)\cdot P(z=P)}{P(y_3=1|z=P)\cdot P(y_1=B, y_2=1|z=P)\cdot P(z=P) + P(y_3=1|z=N)\cdot P(y_1=B, y_2=1|z=N)\cdot P(z=N)} =$$

$$= \frac{1}{\sqrt{2\pi}\cdot 0,0635} \times e^{\frac{-(1-0,84)^2}{2\cdot 0,0635}} \times \frac{1}{5} \times 5/9$$

$$\frac{1}{\sqrt{2\pi}\cdot 0,0635} \times e^{\frac{-(1-0,84)^2}{2\cdot 0,0635}} \times \frac{1}{5} \times \frac{5}{9} + \frac{1}{\sqrt{2\pi}\cdot 0,0292} \times e^{\frac{-(1-0,8975)^2}{2\cdot 0,0292}} \times \frac{1}{4} \times 4/9 = 0,36$$

$x_3 = \begin{pmatrix} B \\ 0 \\ 0,9 \end{pmatrix}$:

$$\frac{P(y_3=0,9|z=P)\cdot P(y_1=B, y_2=0|z=P)\cdot P(z=P)}{P(y_3=0,9|z=P)\cdot P(y_1=B, y_2=0|z=P)\cdot P(z=P) + P(y_3=0,9|z=N)\cdot P(y_1=B, y_2=0|z=N)\cdot P(z=N)} =$$

$$= \frac{\frac{1}{\sqrt{2\pi}\cdot 0,0635} \times e^{\frac{-(0,9-0,84)^2}{2\cdot 0,0635}} \times \frac{1}{5} \times \frac{5}{9}}{\frac{1}{\sqrt{2\pi}\cdot 0,0635} \times e^{\frac{-(0,9-0,84)^2}{2\cdot 0,0635}} \times \frac{1}{5} \times \frac{5}{9} + \frac{1}{\sqrt{2\pi}\cdot 0,0292} \times e^{\frac{-(0,9-0,8975)^2}{2\cdot 0,0292}} \times \frac{3}{4} \times \frac{4}{9}} = 0,28$$

$\theta = 0.3$:

$$f(x|\theta) = \begin{cases} \text{Positiva}; & x = x_1 \vee x = x_2 \\ \text{Negativa}; & x = x_3 \end{cases} \qquad \text{Accuracy} = \frac{\#\text{correct}}{\#\text{predictions}} = 100\%$$

$\theta = 0.5$:

$$f(x|\theta) = \begin{cases} \text{Positiva}; & x = x_1 \\ \text{Negativa}; & x = x_2 \vee x = x_3 \end{cases} \qquad \text{Accuracy} = \frac{\#\text{correct}}{\#\text{predictions}} = 66,(6)\%$$

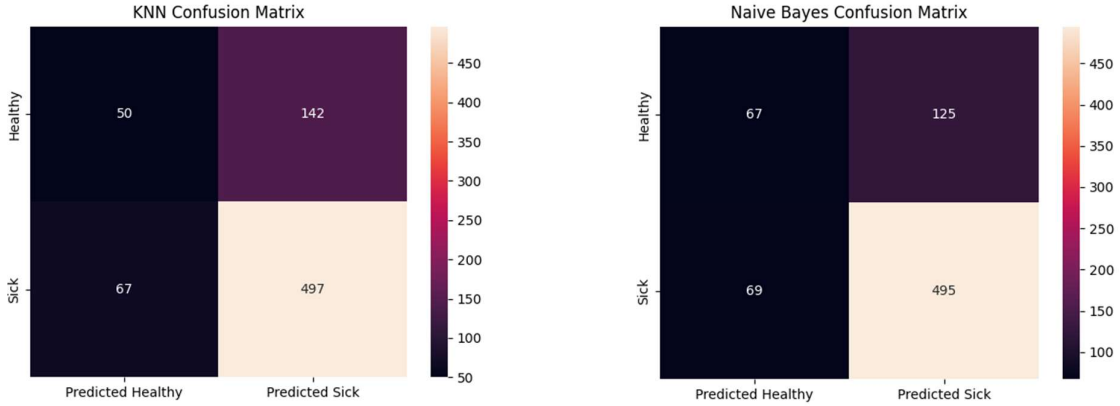
$\theta = 0.7$:

$f(x|\theta) =$ Negativa para $x \in$ conjunto teste; $\text{Accuracy} = \frac{\#\text{correct}}{\#\text{predictions}} = 33,(3)\%$

∴ Podemos concluir que o threshold que otimiza a precisão de teste é $\theta = 0,3$.

II. Programming and critical analysis

5)



6)

Tendo em conta que o p-value = 0.912 da hipótese "kNN é estatisticamente **superior** a Naive Bayes considerando precisão", concluímos que, para um threshold de confiança de 0.1, podemos rejeitar esta hipótese ($0.912 > 0.1$), inferindo, desta forma, que é falsa.

7)

O kNN apenas considera os 5 elementos mais próximos, ao contrário do naive bayes, que tem em conta todo o conjunto. Em adição, o kNN não considerou os pesos dos elementos, diminuindo ainda mais a sua precisão. Para além disso, o naive bayes é um modelo probabilístico que se baseia nos estimadores de máxima verosimilhança.

III. APPENDIX

```
#######         Importing required libraries          #######
import pandas as pd
import seaborn as sns
from scipy import stats
import matplotlib.pyplot as plt
from sklearn import metrics
from sklearn.model_selection import StratifiedKFold
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import confusion_matrix
import numpy as np
from scipy.io.arff import loadarff
import warnings

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


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


#######            Folding and Classifiers            #######
X, y = df.drop('class', axis=1), df['class']
cv = StratifiedKFold(n_splits=10, random_state=0, shuffle=True)
# Creating the classifiers
predictor_kNN = KNeighborsClassifier(weights='uniform', n_neighbors=5,
metric='euclidean')
predictor_nb = GaussianNB()


#######   Running classifier and attesting results    #######
cm_kNN, cm_nb, kNN_acc, nb_acc = [], [], [], []

for train_k, test_k in cv.split(X, y):
    # Getting the training and testing splits
    X_train, X_test = X.iloc[train_k], X.iloc[test_k]
    y_train, y_test = y.iloc[train_k], y.iloc[test_k]
    # Training the classifiers
    predictor_kNN.fit(X_train, y_train)
    predictor_nb.fit(X_train, y_train)
    # Predicting the classes
    y_pred_kNN = predictor_kNN.predict(X_test)
    y_pred_nb = predictor_nb.predict(X_test)
    # Computing the confusion matrices
```

```
    cm_kNN.append(np.array(confusion_matrix(y_test, y_pred_kNN, labels=['0',
'1'])))
    cm_nb.append(np.array(confusion_matrix(y_test, y_pred_nb, labels=['0',
'1'])))
    # Computing the accuracy
    kNN_acc.append(round(metrics.accuracy_score(y_test, y_pred_kNN), 3))
    nb_acc.append(round(metrics.accuracy_score(y_test, y_pred_nb), 3))

cm_kNN = np.sum(cm_kNN, axis=0)
cm_nb = np.sum(cm_nb, axis=0)

# Creating the confusion matrices' plot
confusion_knn = pd.DataFrame(cm_kNN, index=['Healthy', 'Sick', ],
columns=['Predicted Healthy', 'Predicted Sick'])
confusion_nb = pd.DataFrame(cm_nb, index=['Healthy', 'Sick', ],
columns=['Predicted Healthy', 'Predicted Sick'])


#######                    Ex 5                     #######
# KNN confusion matrix
sns.heatmap(confusion_knn, annot=True, fmt='g')
plt.title('KNN Confusion Matrix')
plt.show()

# Naive Bayes confusion matrix
sns.heatmap(confusion_nb, annot=True, fmt='g')
plt.title('Naive Bayes Confusion Matrix')
plt.show()


#######                    Ex 6                     #######
# Performing the t-test
res = stats.ttest_rel(kNN_acc, nb_acc, alternative='greater')
print("knn accuracy: ", kNN_acc)
print("nb accuracy: ", nb_acc)
# Outputting the p-value
print("p1>p2? pval=",np.round(res.pvalue, 3))
```

END