Aprendizagem 2021/22
Homework I – Group 17

I. Pen-and-paper

1) 1) · Cálculo dos priors:

$$P(C) = \begin{cases} \frac{4}{10}, & c=0 \\ \frac{6}{10}, & c=1 \end{cases}$$

· Cálculo dos Posteriors

· $C=0$

$y_1 | c=0 \sim N(\mu, \sigma)$

$$\mu = \overline{y_1} | c=0 = \frac{1}{4}(0,6+0,1+0,2+0,1) = 0,25$$

$$\sigma = \sqrt{V(y_1|c=0)} = \left(\frac{1}{4-1}\left((0,6-0,25)^2+(0,1-0,25)^2+(0,2-0,25)^2+(0,1-0,25)^2\right)\right)^{\frac{1}{2}} = 0,2380$$

$y_1 | C \sim N(0,25 \,;\, 0,2380)$

$y_2 | c=0$

$$P(y_2|c=0) = \begin{cases} \frac{2}{4}, & y_2 = A \\ \frac{1}{4}, & y_2 = B \\ \frac{1}{4}, & y_2 = C \end{cases}$$

$y_3, y_4 | c=0 \sim N(\mu, \Sigma)$

$$\mu = \frac{1}{6}\left(\begin{pmatrix}0,1\\0,3\end{pmatrix} + \begin{pmatrix}0,2\\-0,2\end{pmatrix} + \begin{pmatrix}-0,1\\0,2\end{pmatrix} + \begin{pmatrix}0,5\\0,6\end{pmatrix} + \begin{pmatrix}-0,4\\-0,7\end{pmatrix} + \begin{pmatrix}0,4\\0,3\end{pmatrix}\right) = \begin{pmatrix}0,2\\0,25\end{pmatrix}$$

$$\Sigma = \begin{bmatrix} cov(y_3,y_3) & cov(y_3,y_4) \\ cov(y_4,y_3) & cov(y_4,y_4) \end{bmatrix}, \quad cov(x,y) = \frac{1}{n-1}\sum (x_i-\bar{x})(y_i-\bar{y})$$

$$= \begin{bmatrix} 0,18 & 0,18 \\ 0,18 & 0,25 \end{bmatrix} \qquad \Sigma^{-1} = \begin{pmatrix} 19,841 & -14,286 \\ -14,286 & 14,286 \end{pmatrix} \qquad |\Sigma^{-1}| = 79,3651$$

$$\Sigma^{-1} = \frac{1}{|\Sigma|}\begin{bmatrix} C_{00} & C_{01} \\ C_{10} & C_{11} \end{bmatrix} = \begin{bmatrix} 19,841 & -14,286 \\ -14,286 & 14,286 \end{bmatrix} \;;\; |\Sigma^{-1}| = 19,841 \times 14,286 - 14,286^2 = 79,3651$$

· $C=1$

$y_1 | c=1 \sim N(\mu, \sigma) \qquad \sigma = 0,2881$

$$\mu = \overline{y_1} | c=1 = \frac{1}{6}(0,3-0,1-0,3+0,2+0,4-0,2) = 0,05$$

$$\sigma = \sqrt{V(y_1|c=1)} = \frac{1}{5}\sqrt{(0,3-0,05)^2+(-0,1-0,05)^2+(-0,3-0,05)^2+(0,2-0,05)^2+(0,4-0,05)^2+(-0,2-0,05)^2}$$

$$= 0,2881$$

$y_1|c=1 \sim N(0,05 \,;\, 0,2881)$

$y_2 | c=1$

$$P(y_2|c=1) = \begin{cases} \frac{1}{6}, & y_2 = A \\ \frac{2}{6}, & y_2 = B \\ \frac{3}{6}, & y_2 = C \end{cases}$$

$y_3, y_4 | c=1$

$y_3, y_4 \sim N(\mu, \Sigma)$

$$\mu = \frac{1}{6}\left(\begin{pmatrix}0,1\\0,3\end{pmatrix} + \begin{pmatrix}0,2\\-0,2\end{pmatrix} + \begin{pmatrix}-0,1\\0,2\end{pmatrix} + \begin{pmatrix}0,5\\0,6\end{pmatrix} + \begin{pmatrix}-0,4\\-0,7\end{pmatrix} + \begin{pmatrix}0,4\\0,3\end{pmatrix}\right) = \begin{pmatrix}0,1167\\0,0833\end{pmatrix}$$

Aprendizagem 2021/22

Homework I – Group 17

$$\Sigma = \begin{bmatrix} cov(y_3, y_3) & cov(y_3, y_4) \\ cov(y_4, y_3) & cov(y_4, y_4) \end{bmatrix}, \quad cov(x,y) = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y})$$

$$= \begin{bmatrix} 0{,}10967 & 0{,}12233 \\ 0{,}12233 & 0{,}21367 \end{bmatrix}$$

$$\Sigma^{-1} = \frac{1}{|\Sigma|} \begin{bmatrix} C_{00} & C_{01} \\ C_{10} & C_{11} \end{bmatrix} = \begin{bmatrix} 25{,}2362 & -14{,}449 \\ -14{,}449 & 12{,}953 \end{bmatrix}; \quad |\Sigma^{-1}| = 118{,}11$$

- Posteriors

$$P(c) \cdot P(x \mid C=c) = \left(\prod_{i=1}^{4} P(y_i \mid c=c) \right) \cdot P(c)$$

$P(x=x_1 \mid C=0) \cdot P(C=0) = 0{,}15728$
$P(x=x_1 \mid C=1) \cdot P(C=1) = 0{,}02712$

$P(x=x_2 \mid C=0) \cdot P(C=0) = 0{,}06326$
$P(x=x_2 \mid C=1) \cdot P(C=1) = 0{,}26104$

$P(x=x_3 \mid C=0) \cdot P(C=0) = 0{,}23172$
$P(x=x_3 \mid C=1) \cdot P(C=1) = 0{,}07347$

$P(x=x_4 \mid C=0) \cdot P(C=0) = 0{,}07041$
$P(x=x_4 \mid C=1) \cdot P(C=1) = 0{,}08310$

$P(x=x_5 \mid C=0) \cdot P(C=0) = 0{,}19254$
$P(x=x_5 \mid C=1) \cdot P(C=1) = 0{,}22937$

$P(x=x_6 \mid C=0) \cdot P(C=0) = 0{,}01898$
$P(x=x_6 \mid C=1) \cdot P(C=1) = 0{,}24307$

$P(x=x_7 \mid C=0) \cdot P(C=0) = 0{,}00820$
$P(x=x_7 \mid C=1) \cdot P(C=1) = 0{,}12068$

$P(x=x_8 \mid C=0) \cdot P(C=0) = 0{,}17785$
$P(x=x_8 \mid C=1) \cdot P(C=1) = 0{,}20335$

$P(x=x_9 \mid C=0) \cdot P(C=0) = 0{,}05976$
$P(x=x_9 \mid C=1) \cdot P(C=1) = 0{,}02569$

$P(x=x_{10} \mid C=0) \cdot P(C=0) = 0{,}03031$
$P(x=x_{10} \mid C=1) \cdot P(C=1) = 0{,}32083$

2) 2) testes:

$P(C=0 \mid x_{new}=x_1) = 0{,}835$
$P(C=1 \mid x_{new}=x_1) = 0{,}165$

$P(C=0 \mid x_{new}=x_2) = 0{,}195$
$P(C=1 \mid x_{new}=x_2) = 0{,}805$

$P(C=0 \mid x_{new}=x_3) = 0{,}759$
$P(C=1 \mid x_{new}=x_3) = 0{,}241$

$P(C=0 \mid x_{new}=x_4) = 0{,}459$
$P(C=1 \mid x_{new}=x_4) = 0{,}541$

$P(C=0 \mid x_{new}=x_5) = 0{,}456$
$P(C=1 \mid x_{new}=x_5) = 0{,}544$

$P(C=0 \mid x_{new}=x_6) = 0{,}072$
$P(C=1 \mid x_{new}=x_6) = 0{,}928$

$P(C=0 \mid x_{new}=x_7) = 0{,}064$
$P(C=1 \mid x_{new}=x_7) = 0{,}936$

$P(C=0 \mid x_{new}=x_8) = 0{,}467$
$P(C=1 \mid x_{new}=x_8) = 0{,}533$

$P(C=0 \mid x_{new}=x_9) = 0{,}699$
$P(C=1 \mid x_{new}=x_9) = 0{,}301$

$P(C=0 \mid x_{new}=x_{10}) = 0{,}086$
$P(C=1 \mid x_{new}=x_{10}) = 0{,}914$

Confusion Matrix

		Predicted	
		C=0	C=1
True	C=0	2	2
	C=1	1	5

Aprendizagem 2021/22

Homework I – Group 17

3) 3)

$$F1(c=0) = \left(\frac{R(c=0)^{-1} + P(c=0)^{-1}}{2}\right)^{-1} = \left(\frac{2 + \frac{3}{2}}{2}\right)^{-1} = \left(\frac{7}{4}\right)^{-1} = \frac{4}{7}$$

$$R(c=0) = \frac{TP}{TP+FN} = 1/2$$

$$P(c=0) = \frac{TP}{TP+FP} = 2/3$$

$$F1(c=1) = \left(\frac{R(c=1)^{-1} + P(c=1)^{-1}}{2}\right)^{-1} = \left(\frac{\frac{6}{5} + \frac{7}{5}}{2}\right)^{-1} = \left(\frac{13}{10}\right)^{-1} = \frac{10}{13}$$

$$R(c=1) = \frac{TP}{TP+FN} = \frac{5}{6}$$

$$P(c=1) = \frac{TP}{TP+FP} = \frac{5}{7}$$

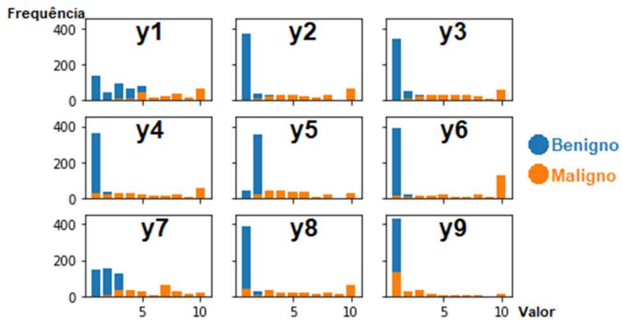
4) 4)

thresholds

Var.	Verdade	P(c=1)	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0
x_1	0	0,1650	1	1	0	0	0	0	0	0	0	0	0
x_2	0	0,8049	1	1	1	1	1	1	1	1	0	0	0
x_3	0	0,2408	1	1	1	0	0	0	0	0	0	0	0
x_4	0	0,5413	1	1	1	1	1	1	0	0	0	0	0
x_5	1	0,5436	1	1	1	1	1	1	0	0	0	0	0
x_6	1	0,9276	1	1	1	1	1	1	1	1	1	1	0
x_7	1	0,9364	1	1	1	1	1	1	1	1	1	1	0
x_8	1	0,5334	1	1	1	1	1	1	0	0	0	0	0
x_9	1	0,3006	1	1	1	1	0	0	0	0	0	0	0
x_{10}	1	0,9137	1	1	1	1	1	1	1	1	1	1	0
Accuracy:			$\frac{3}{5}$	$\frac{3}{5}$	$\frac{7}{10}$	$\frac{4}{5}$	$\frac{7}{10}$	$\frac{7}{10}$	$\frac{2}{5}$	$\frac{2}{5}$	$\frac{3}{10}$	$\frac{3}{10}$	$\frac{2}{5}$

O valor de threshold que maximiza a accuracy do classificador é 0,30 pois ao verificar empiricamente, é para este valor que obtemos o maior número de previsões certas.

II. Programming and critical analysis

5) 

6) Accuracies:
k = 3: 0.9707, k = 5: 0.9751, k = 7: 0.9678
Logo, conclui-se que k = 5 é o menos suscetível a overfitting.

7) Naïve Bayes accuracy: 0.9047
Para testar a hipótese “kNN é estatisticamente superior ao Naïve Bayes” tomamos como H0 a afirmação “Naïve Bayes é estatisticamente superior ao kNN” e como H1 a negação de H0. Ao fazer um t-test unilateral, verificamos que é possível rejeitar H0 para p-value 1.46e-5, que é muito baixo.

8) As diferenças de performance entre o kNN e o Naïve Bayes podem ser explicadas pelo facto de as class conditional distributions serem skewed, especialmente as de C = 0 (positivamente), e, para

além disso, por não se assumir independência entre os atributos, o que leva a que o kNN tenha melhores resultados.

III. APPENDIX

```
import scipy.io.arff; import matplotlib.pyplot as plt; import numpy as np
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import KFold
from sklearn.naive_bayes import MultinomialNB
from scipy.stats import ttest_ind
# ficheiro sem observacoes com entradas "?"
FILE_NAME, SEED, genmat = "breast.w.new.arff", 17, lambda:[[0 for i in range(10)] for i in range(9)]

data, meta = scipy.io.arff.loadarff(FILE_NAME)

#question 5
benign_m, malign_m = genmat(), genmat()

for entry in data:
    if entry['Class'] == b'benign': matrix = benign_m
    else: matrix = malign_m
    for i in range(9): matrix[i][int(entry[i]) - 1] += 1

fig, ax = plt.subplots(3,3, sharex=True, sharey=True)

for n in range(9):
    ax[n // 3, n % 3].bar([el for el in range(1,11)], benign_m[n])
    ax[n // 3, n % 3].bar([el for el in range(1,11)], malign_m[n])

plt.show()

#questions 6/7
knn_classifiers = dict()
# separate attributes from classes
targets = list(el["Class"] for el in data)
training_data = np.array(list(list(el[i] for i in range(9)) for el in data))
k_fold = KFold(n_splits=10, shuffle=True, random_state=SEED)
for n in (3, 5, 7): knn_classifiers[n] = KNeighborsClassifier(n_neighbors=n, p=2, weights="uniform")

naive_bayes_classifier, folds, nb_folds, i = MultinomialNB(), [], [], 1
for train, test in k_fold.split(training_data):
    for n in (3, 5, 7):
        knn_classifiers[n].fit(np.array([training_data[i] for i in train]),\
                               np.array([targets[i] for i in train]))
        acc = knn_classifiers[n].score(np.array([training_data[i] for i in test]),\
                               np.array([targets[i] for i in test]), sample_weight=None)
        folds.append({ "fold" : i, "n" : n, acc" : acc })

    naive_bayes_classifier.fit(np.array([training_data[i] for i in train]),\
                               np.array([targets[i] for i in train]))
    nb_acc = naive_bayes_classifier.score(np.array([training_data[i] for i in test]),\
                               np.array([targets[i] for i in test]), sample_weight=None)
    nb_folds.append({ "fold" : i, "acc" : nb_accs})
    i += 1

accs, nb_accs, avg_accs, avg_nb_acc = { 3 : [], 5 : [], 7 : [] }, [], { 3 : 0, 5 : 0, 7 : 0 }, 0
for fold in folds: accs[fold['n']].append(fold['acc'])
for nb_fold in nb_folds: nb_accs.append(nb_fold["acc"])
for n in accs: avg_accs[n] = sum(accs[n]) / 10

avg_nb_acc = sum(nb_accs) / 10
knn_3_acc = accs[3]
statistic, pvalue = ttest_ind(knn_3_acc, nb_accs, alternative="less") #hypothesis test
```

END