Aprendizagem 2022/23

Homework I - Group 22

I. Pen-and-paper

1) By building a table like the following, it's easy to compute the distances between observations. The blue-highlighted cells represent the 5 closest neighbors. The distances were then weighed to make predictions ($\frac{1}{d(x_1,x_2)}$), e.g., for $(A\ 0)\ \rightarrow\ \frac{2}{3} + 2 > \frac{2}{3} + \frac{2}{3} + \frac{2}{3}$ which means the prediction is going to be **positive**.

x_1/x_2	$(A\ 0)$	(B 1)	(A 1)	(A 0)	(B 0)	(B 0)	(A 1)	(B 1)	Prediction
$(A\ 0)$	X	$\frac{5}{2}$	$\frac{3}{2}$	$\frac{1}{2}$	$\frac{3}{2}$	$\frac{3}{2}$	$\frac{3}{2}$	$\frac{5}{2}$	+
(B 1)	$\frac{5}{2}$	X	$\frac{3}{2}$	$\frac{5}{2}$	$\frac{3}{2}$	$\frac{3}{2}$	$\frac{3}{2}$	$\frac{1}{2}$	−
(A 1)	$\frac{3}{2}$	$\frac{3}{2}$	X	$\frac{3}{2}$	$\frac{5}{2}$	$\frac{5}{2}$	$\frac{1}{2}$	$\frac{3}{2}$	−
(A 0)	$\frac{1}{2}$	$\frac{5}{2}$	$\frac{3}{2}$	X	$\frac{3}{2}$	$\frac{3}{2}$	$\frac{3}{2}$	$\frac{5}{2}$	+
(B 0)	$\frac{3}{2}$	$\frac{3}{2}$	$\frac{5}{2}$	$\frac{3}{2}$	X	$\frac{1}{2}$	$\frac{5}{2}$	$\frac{3}{2}$	−
(B 0)	$\frac{3}{2}$	$\frac{3}{2}$	$\frac{5}{2}$	$\frac{3}{2}$	$\frac{1}{2}$	X	$\frac{5}{2}$	$\frac{3}{2}$	−
(A 1)	$\frac{3}{2}$	$\frac{3}{2}$	$\frac{1}{2}$	$\frac{3}{2}$	$\frac{5}{2}$	$\frac{5}{2}$	X	$\frac{3}{2}$	+
(B 1)	$\frac{5}{2}$	$\frac{1}{2}$	$\frac{3}{2}$	$\frac{5}{2}$	$\frac{3}{2}$	$\frac{3}{2}$	$\frac{3}{2}$	X	+

Finally, recall can be computed.

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

2) To a learn a Bayesian classifier we need to find the parameters to compute $P(class\ =\ Positive|X)$ as a function of Y_1, Y_2 and Y_3.

We start by building a table with all of the training data.

	Y_1	Y_2	Y_3	**Class**
X_1	A	0	1,2	Positive
X_2	B	1	0,8	Positive
X_3	A	1	0,5	Positive
X_4	A	0	0,9	Positive
X_5	B	0	0,8	Positive
X_6	B	0	1	Negative
X_7	B	0	0,9	Negative
X_8	A	1	1,2	Negative
X_9	B	1	0,8	Negative

Using Bayes' theorem, we get:

$$P(class\ =\ Positive|X)\ =\ \frac{P(X|class=Positive)P(class=Positive)}{P(X)}$$

To compute P(X) we use the law of total probability:

$$P(X)\ =\ P(X|class\ =\ Positive)P(class\ =\ Positive)\ +\ P(X|class\ =\ Negative)P(class\ =\ Negative)$$

And we can compute $P(X|class\ =\ Positive)$ and $P(X|class\ =\ Negative)$ by having in account that Y_1 and Y_2 are dependent and $\{y1,\ y2\}$ and $\{y3\}$ variable sets are independent and equally important.

As such:

$$P(X|class = Positive) = P(Y_1, Y_2|class = Positive)P(Y_3|class = Positive)$$

$$P(X|class = Negative) = P(Y_1, Y_2|class = Negative)P(Y_3|class = Negative)$$

The final expression of $P(class = Positive|X)$ as a function of Y_1, Y_2 and Y_3 goes as follows:

$$P(class = Positive|X) =$$

$$= \frac{P(Y_1,Y_2|class=Positive)P(Y_3|class=Positive)P(class=Positive)}{P(Y_1,Y_2|class=Positive)P(Y_3|class=Positive)P(class=Positive) + P(Y_1,Y_2|class=Negative)P(Y_3|class=Negative)P(class=Negative)}$$

We can now compute all the parameters for $P(class = Positive|X)$ as a function of Y_1, Y_2 and Y_3.

$P(class = Positive) = \frac{5}{9}$

$P(class = Negative) = \frac{4}{9}$

$P(Y_1, Y_2|class = Positive)$**:**

$$P(Y_1 = A, Y_2 = 0|class = Positive) = \frac{2}{5}$$

$$P(Y_1 = A, Y_2 = 1|class = Positive) = \frac{1}{5}$$

$$P(Y_1 = B, Y_2 = 0|class = Positive) = \frac{1}{5}$$

$$P(Y_1 = B, Y_2 = 1|class = Positive) = \frac{1}{5}$$

$P(Y_1, Y_2|class = Negative)$**:**

$$P(Y_1 = A, Y_2 = 0|class = Negative) = 0$$

$$P(Y_1 = A, Y_2 = 1|class = Negative) = \frac{1}{4}$$

$$P(Y_1 = B, Y_2 = 0|class = Negative) = \frac{2}{4}$$

$$P(Y_1 = B, Y_2 = 1|class = Negative) = \frac{1}{4}$$

As Y_3 is normally distributed $P(Y_3|class) \sim N(Y_3, \mu, \sigma^2)$ and $P(Y_3) \sim N(Y_3, \mu, \sigma^2)$

$$\mu = \frac{\Sigma x_i}{n} \qquad \sigma^2 = \frac{\Sigma(x_i - \mu)^2}{n-1}$$

$$N(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

$P(Y_3|class\ =\ Positive)$**:**

$$\mu\ =\ \frac{1,2+0.8+0.5+0.9+0.8}{5}\ =\ 0.84$$

$$\sigma^2\ =\ \frac{(1.2-0.84)^2+(0.8-0.84)^2+(0.5-0.84)^2+(0.9-0.84)^2+(0.8-0.84)^2}{4}\ \approx\ 0.063$$

$$P(Y_3|class\ =\ Positive)\ =\ \frac{1}{\sqrt{2\pi\cdot 0.063}}exp\left(\frac{-(Y_3-0.84)^2}{2\times 0.063}\right)$$

$P(Y_3|class\ =\ Negative)$**:**

$$\mu\ =\ \frac{1,2+0.8+0.5+0.9+0.8}{5}\ =\ 0.975$$

$$\sigma^2\ =\ \frac{(1.2-0.84)^2+(0.8-0.84)^2+(0.5-0.84)^2+(0.9-0.84)^2+(0.8-0.84)^2}{4}\ \approx\ 0.0292$$

$$P(Y_3|class\ =\ Negative)\ =\ \frac{1}{\sqrt{2\pi\cdot 0.0292}}exp\left(\frac{-(Y_3-0.975)^2}{2\times 0.0292}\right)$$

3) According to exercise 2

$$P(class\ =\ Positive|X)\ =$$

$$=\frac{P(Y_1,Y_2|class=Positive)P(Y_3|class=Positive)P(class=Positive)}{P(Y_1,Y_2|class=Positive)P(Y_3|class=Positive)P(class=Positive)\ +\ P(Y_1,Y_2|class=Negative)P(Y_3|class=Negative)P(class=Negative)}$$

$$P(class\ =\ Positive)\ \ =\frac{5}{9}$$

$$P(class\ =\ Negative)\ \ =\frac{4}{9}$$

P(Positive$|o_1$):

$$P(Y_1\ =\ A, Y_2\ =\ 1|class\ =\ Positive)\ =\frac{1}{5}$$

$$P(Y_1\ =\ A, Y_2\ =\ 1|class\ =\ Negative)\ =\frac{1}{4}$$

$$P(Y_3\ =\ 0.8|class\ =\ Positive)\ =\frac{1}{\sqrt{2\pi\cdot 0.063}}exp\left(\frac{-(0.8-0.84)^2}{2\times 0.063}\right)=\ 1.569$$

$$P(Y_3\ =\ 0.8|class\ =\ Negative)\ =\frac{1}{\sqrt{2\pi\cdot 0.0292}}exp\left(\frac{-(0.8-0.975)^2}{2\times 0.0292}\right)=\ 1.382$$

$$P(Positive|o_1)\ \ =\frac{\frac{1}{5}\times 1.569\times\frac{5}{9}}{\frac{1}{5}\times 1.569\times\frac{5}{9}+\frac{1}{4}\times 1.382\times\frac{4}{9}}\ \approx\ 0.5317$$

$P(\text{Positive}|o_2)$:

$$P(Y_1 = B, Y_2 = 1|class = Positive) = \frac{1}{5}$$

$$P(Y_1 = B, Y_2 = 1|class = Negative) = \frac{1}{4}$$

$$P(Y_3 = 1|class = Positive) = \frac{1}{\sqrt{2\pi \cdot 0.063}} exp\left(\frac{-(1-0.84)^2}{2\times 0.063}\right) = 1.297$$

$$P(Y_3 = 1|class = Negative) = \frac{1}{\sqrt{2\pi \cdot 0.0292}} exp\left(\frac{-(1-0.975)^2}{2\times 0.0292}\right) = 2.311$$

$$P(Positive|o_2) = \frac{\frac{1}{5}\times 1.297 \times \frac{5}{9}}{\frac{1}{5}\times 1.297 \times \frac{5}{9} + \frac{1}{4}\times 2.311 \times \frac{4}{9}} \approx 0.3595$$

$P(\text{Positive}|o_3)$:

$$P(Y_1 = B, Y_2 = 0|class = Positive) = \frac{1}{5}$$

$$P(Y_1 = B, Y_2 = 0|class = Negative) = \frac{2}{4}$$

$$P(Y_3 = 0.9|class = Positive) = \frac{1}{\sqrt{2\pi \cdot 0.063}} exp\left(\frac{-(0.9-0.84)^2}{2\times 0.063}\right) = 1.545$$

$$P(Y_3 = 0.9|class = Negative) = \frac{1}{\sqrt{2\pi \cdot 0.0292}} exp\left(\frac{-(0.9-0.975)^2}{2\times 0.0292}\right) = 2.121$$

$$P(Positive|o_3) = \frac{\frac{1}{5}\times 1.545 \times \frac{5}{9}}{\frac{1}{5}\times 1.545 \times \frac{5}{9} + \frac{2}{4}\times 2.121 \times \frac{4}{9}} \approx 0.2670$$

To check if we need to normalize the variables, we also calculate $P(\text{Negative}|\mathbf{x})$

$$P(Negative|o_1) = \frac{\frac{1}{4}\times 1.382 \times \frac{4}{9}}{\frac{1}{5}\times 1.569 \times \frac{5}{9} + \frac{1}{4}\times 1.382 \times \frac{4}{9}} \approx 0.4683$$

$$P(Negative|o_2) = \frac{\frac{1}{4}\times 2.311 \times \frac{4}{9}}{\frac{1}{5}\times 1.297 \times \frac{5}{9} + \frac{1}{4}\times 2.311 \times \frac{4}{9}} \approx 0.6405$$

$$P(Negative|o_3) = \frac{\frac{2}{4}\times 2.121 \times \frac{4}{9}}{\frac{1}{5}\times 1.545 \times \frac{5}{9} + \frac{2}{4}\times 2.121 \times \frac{4}{9}} \approx 0.7330$$

As $P(Positive|x) + P(Negative|x) \approx 1$ for any $x \in \{o_1, o_2, o_3\}$, we won't be normalizing $P(Positive|x)$, even though that would be the most correct way of calculating the probability.

4) Accuracy(θ = 0.3) = 1

Accuracy(θ = 0.5) = ⅔

Accuracy(θ = 0.7) = ⅓

θ = 0.3 optimizes testing accuracy.

II. Programming and critical analysis

5) Code solution is provided in Appendix (1) and the plotted confusion matrices are provided in Appendix (2). The data was normalized, which greatly helped kNN.

6) To test the given hypothesis, we must first formulate the Null Hypothesis.

Null Hypothesis: kNN is not statistically superior to Naïve Bayes.

A paired single-tailed t-test based on the 10-fold testing estimates was done, the code and output of which are provided in Appendix (3).

The resulting p-value was approximately 0.001317.

As such, the Null hypothesis is rejected at 1%, confirming the statistical superiority of kNN comparatively to Naïve Bayes.

7) Naïve Bayes assumes conditional independence between variables, which isn't always the case. Additionally, due to the moderate data size, the pdf approximations in Naïve Bayes might suffer.

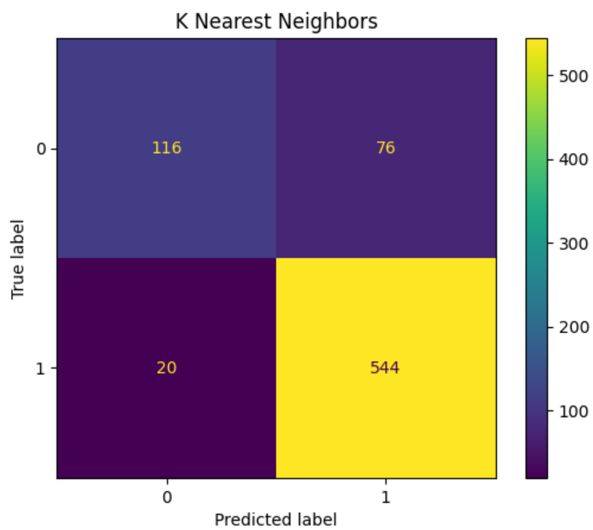
III. APPENDIX

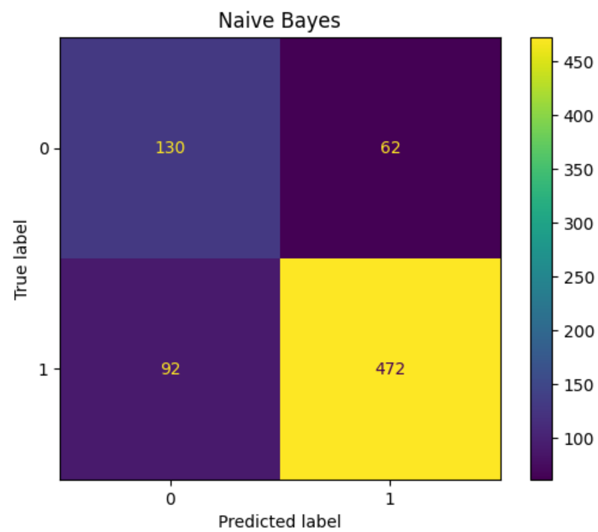
1)

```
import pandas as pd
from scipy.io.arff import loadarff

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

X = df.drop('class', axis=1)
y = df['class']

from sklearn.model_selection import StratifiedKFold
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay
from sklearn.preprocessing import StandardScaler
import numpy as np

# Stratified K-Fold
folds = StratifiedKFold(n_splits=10, shuffle=True, random_state=0)

# KNN and Naive Bayes
knn = KNeighborsClassifier(n_neighbors=5, weights='uniform', metric='euclidean')
nb = GaussianNB()

# initialize arrays to store the confusion matrices
cm_knn = []
cm_nb = []

# initialize lists to store the accuracy of each fold
acc_knn = []
acc_nb = []
```

```python
# initialize normalization scaler
scaler = StandardScaler()

# iterate per fold
for train_k, test_k in folds.split(X, y):
    # split the data
    X_train, X_test = X.iloc[train_k], X.iloc[test_k]
    y_train, y_test = y.iloc[train_k], y.iloc[test_k]
    # normalize the data
    X_train = scaler.fit_transform(X_train)
    X_test = scaler.transform(X_test)
    # train and assess
    knn.fit(X_train, y_train)
    nb.fit(X_train, y_train)
    # add prediction to cumulative confusion matrix
    cm_knn.append(confusion_matrix(y_test, knn.predict(X_test)))
    cm_nb.append(confusion_matrix(y_test, nb.predict(X_test)))
    # add accuracy to list
    acc_knn.append(knn.score(X_test, y_test))
    acc_nb.append(nb.score(X_test, y_test))


# plot the confusion matrices
disp_knn = ConfusionMatrixDisplay(sum(cm_knn))
disp_knn.plot()
disp_knn.ax_.set_title("K Nearest Neighbors")
disp_nb = ConfusionMatrixDisplay(sum(cm_nb))
disp_nb.plot()
disp_nb.ax_.set_title("Naive Bayes")
```

2)





3)

```python
# t-test to test the hypothesis
from scipy.stats import ttest_rel
t, p = ttest_rel(acc_knn, acc_nb, alternative='less')
print("knn < naive bayes: p =", p)
t, p = ttest_rel(acc_knn, acc_nb, alternative='greater')
print("knn > naive bayes: p =", p)
t, p = ttest_rel(acc_knn, acc_nb, alternative='two-sided')
print("knn == naive bayes: p =", p)
```

Output:

```
knn < naive bayes: p = 0.9986831821715092
knn > naive bayes: p = 0.0013168178284908261
knn == naive bayes: p = 0.0026336356569816522
```