TÉCNICO LISBOA

Machine Learning

LEIC IST-UL

RELATÓRIO - HOMEWORK 1

Grupo 10:

Gabriel Ferreira	107030
Irell Zane	107161

2024/2025 – 1st Semester, P1

Part I: Pen and paper

1. F1-measure of a kNN.

	P				N			
	x1	x2	x3	x4	x5	x6	x7	x8
x1	-	2	1	0	1	1	1	2
x2	2	-	1	2	1	1	1	0
x3	1	1	-	1	2	2	0	1
x4	0	2	1	-	1	1	1	2
x5	1	1	2	1	-	0	2	1
x6	1	1	2	1	0	-	2	1
x7	1	1	0	1	2	2	-	1
x8	2	0	1	2	1	1	1	-

Table 1: Hamming distance between observations

Thus with $k = 5$, and a leave-one-out evaluation schema, we use the closest 5 observations for each, excluding itself, to calculate the estimate using a weighted mode like so:

$$f(x_{new}) \leftarrow \underset{c \in \{P,N\}}{\operatorname{argmax}} \sum^{i} w_i \cdot \delta(c, f(x_i))$$

$$w_i = \begin{cases} \frac{1}{d(x_new, x_i)} & \text{if } x_{new} \neq x_i \\ 1 & else \end{cases}$$

In this case, because the relevant observations all have distances of either 0 or 1, the weight of each is the same:

	P				N						
	x1	x2	x3	x4	x5	x6	x7	x8	P	N	$f(x_{new})$
x1	-	-	1	0	1	1	1	-	2	3	N
x2	-	-	1	-	1	1	1	0	1	4	N
x3	1	1	-	1	-	-	0	1	3	2	P
x4	0	-	1	-	1	1	1	-	2	3	N
x5	1	1	-	1	-	0	-	1	3	2	P
x6	1	1	-	1	0	-	-	1	3	2	P
x7	1	1	0	1	-	-	-	1	4	1	P
x8	-	0	1	-	1	1	1	-	2	3	N

Table 2: leave-one-out evaluation kNN classifications

Now the confusion matrix:

	P	N
P	1	3
N	3	1

To calculate the F1-Measure we now need Precision and Recall:

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{1}{4}$$

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{1}{4}$$

I.1 Solution:

$$\text{F1 Score} = \frac{2}{\frac{1}{P} + \frac{1}{R}} = \frac{2}{4+4} = \mathbf{\frac{1}{4}}$$

2. An example of a distance and k that will improve the F1-Measure by three fold is the following:

I.2 Solution:

$$d(x_1, x_2) = 2 \cdot d_{y_1}(x_1, x_2) + d_{y_2}(x_1, x_2)$$

$$k = 3$$

Where $d_{y_j}(x_1, x_2)$ is the Hamming distance between x_1 and x_2 considering only the variable y_j.

To demonstrate the same process as previous but with the new distance measure and k value:

	P				N			
	x1	x2	x3	x4	x5	x6	x7	x8
x1	-	3	1	0	2	2	1	2
x2	3	-	2	3	1	1	2	0
x3	1	2	-	1	3	3	0	2
x4	0	3	1	-	2	2	1	2
x5	2	1	3	2	-	0	3	1
x6	2	1	3	2	0	-	3	1
x7	1	2	0	1	3	3	-	2
x8	3	0	2	3	1	1	2	-

Table 3: New distance between observations

			P				N				
	x1	x2	x3	x4	x5	x6	x7	x8	P	N	$f(x_{new})$
x1	-	-	1	0	-	-	1	-	2	1	P
x2	-	-	-	-	1	1	-	0	0	3	N
x3	1	-	-	1	-	-	0	-	2	1	P
x4	0	-	1	-	-	-	1	-	2	1	P
x5	-	1	-	-	-	0	-	1	1	2	N
x6	-	1	-	-	0	-	-	1	1	2	N
x7	1	-	0	1	-	-	-	-	3	0	P
x8	-	0	-	-	1	1	-	-	1	2	N

Table 4: leave-one-out evaluation with new metric

This metric performs better in this data set as can be seen in the confusion matrix:

	P	N
P	3	1
N	1	3

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{3}{4}$$

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{3}{4}$$

$$\text{F1 Score} = \frac{2}{\frac{1}{P} + \frac{1}{R}} = \frac{2}{\frac{4}{3} + \frac{4}{3}} = \mathbf{\frac{3}{4}}$$

3. (a) **Dataset**:

x	y_1	y_2	y_3	Class
1	A	0	1.1	P
2	B	1	0.8	P
3	A	1	0.5	P
4	A	0	0.9	P
5	B	0	1.0	N
6	B	0	0.9	N
7	A	1	1.2	N
8	B	1	0.9	N
9	B	0	0.8	P

Table 5: Observed Values

(b) **Priors**:

- For class Positive (P):

$$p(P) = \frac{5}{9}$$

- For class Negative (N):

$$p(N) = \frac{4}{9}$$

(c) **Class-conditional Probabilities**:

- Calculate the probabilities of the variable set $\{y_1, y_2\}$ given each class.
- The values are calculated as follows:

$$\begin{aligned}
p(y_1 = A, y_2 = 0) &= \frac{2}{9} \\
p(y_1 = A, y_2 = 1) &= \frac{2}{9} \\
p(y_1 = B, y_2 = 0) &= \frac{2}{9} \\
p(y_1 = B, y_2 = 1) &= \frac{3}{9} \\
p(y_1 = A, y_2 = 0|P) &= \frac{2}{5} \\
p(y_1 = A, y_2 = 1|P) &= \frac{1}{5} \\
p(y_1 = B, y_2 = 0|P) &= \frac{1}{5} \\
p(y_1 = B, y_2 = 1|P) &= \frac{1}{5} \\
p(y_1 = A, y_2 = 0|N) &= 0 \\
p(y_1 = A, y_2 = 1|N) &= \frac{1}{4} \\
p(y_1 = B, y_2 = 0|N) &= \frac{2}{4} \\
p(y_1 = B, y_2 = 1|N) &= \frac{1}{4}
\end{aligned}$$

(d) **Mean and Standard Deviation of** y_3:

- For all observations:

$$\begin{aligned}
\mu_{y_3} &= \frac{1.1 + 0.8 + 0.5 + 0.9 + 0.8 + 1.0 + 0.9 + 1.2 + 0.9}{9} \\
&= 0.9
\end{aligned}$$

$$\begin{aligned}
\sigma_{y_3} &= \sqrt{\frac{(1.1 - 0.9)^2 + (0.8 - 0.9)^2 + (0.5 - 0.9)^2 + (0.9 - 0.9)^2 ...}{8}} \\
&\approx 0.2
\end{aligned}$$

- For Positive (P):

$$\mu_{y_3,P} = \frac{1.1 + 0.8 + 0.5 + 0.9 + 0.8}{5} = 0.82$$
$$\sigma_{y_3,P} = \sqrt{\frac{(1.1 - 0.82)^2 + (0.8 - 0.82)^2 + (0.5 - 0.82)^2 + ...}{4}}$$
$$\approx 0.217$$

- For Negative (N):

$$\mu_{y_3,N} = \frac{1.0 + 0.9 + 1.2 + 0.9}{4} = 1.0$$
$$\sigma_{y_3,N} = \sqrt{\frac{(1.0 - 1.0)^2 + (0.9 - 1.0)^2 + (1.2 - 1.0)^2 + (0.9 - 1.0)^2}{3}}$$
$$\approx 0.1414$$

(e) **Prediction of Class**:

- To predict the class of a new observation (y_1, y_2, y_3), we calculate the probability for both classes (Positive and Negative) and choose the class with the higher probability:

$$\text{Predicted Class} = \underset{h}{\text{argmax}}\, p(h|y_1, y_2, y_3)$$

$$\text{where } p(h|y_1, y_2, y_3) = \frac{p(y_1, y_2, y_3|h) \cdot p(h)}{p(y_1, y_2, y_3)}$$
$$= \frac{p(y_1, y_2|h) \cdot p(y_3|h) \cdot p(h)}{p(y_1, y_2) \cdot p(y_3)}$$
$$\text{where } p(y_3|h) = \frac{1}{\sigma_h\sqrt{2\pi}} \exp\left(-\frac{(y_3 - \mu_h)^2}{2\sigma_h^2}\right)$$
$$\text{and } p(y_3) = \frac{1}{0.2\sqrt{2\pi}} \exp\left(-\frac{(y_3 - 0.9)^2}{2 \cdot 0.2^2}\right)$$

4. Under a MAP assumption we do not need to calculate the denominator, thus:

$$\text{Predicted Class} = \underset{h}{\text{argmax}}\{p(y_1, y_2|h) \cdot p(y_3|h) \cdot p(h)\}$$

(a) For observation (A, 1, 0.8):

For class P:

$$p(y_3 = 0.8|P) = \frac{1}{0.217\sqrt{2\pi}} \exp\left(-\frac{(0.8-0.82)^2}{2 \cdot 0.217^2}\right)$$

$$= 1.83$$

$$p(y_3 = 0.8) = \frac{1}{0.2\sqrt{2\pi}} \exp\left(-\frac{(0.8-0.9)^2}{2 \cdot 0.2^2}\right)$$

$$= 1.76$$

$$p(P) \cdot p(y_1 = A, y_2 = 1, y_3 = 0.8|P) = \frac{5}{9} \cdot \frac{1}{5} \cdot 1.83$$

$$\approx 0.203$$

For class N:

$$p(y_3 = 0.8|N) = \frac{1}{0.1414\sqrt{2\pi}} \exp\left(-\frac{(0.8-1.0)^2}{2 \cdot 0.1414^2}\right)$$

$$= 1.038$$

$$p(N) \cdot p(y_1 = A, y_2 = 1, y_3 = 0.8|N) = \frac{4}{9} \cdot \frac{1}{4} \cdot 1.038$$

$$\approx 0.115$$

Since $0.203 > 0.115$, the predicted class is P.

(b) For observation (B, 1, 1):

For class P:

$$p(y_3 = 1|P) = \frac{1}{0.217\sqrt{2\pi}} \exp\left(-\frac{(1-0.82)^2}{2 \cdot 0.217^2}\right)$$

$$= 1.304$$

$$p(y_3 = 1) = \frac{1}{0.2\sqrt{2\pi}} \exp\left(-\frac{(1-0.9)^2}{2 \cdot 0.2^2}\right)$$

$$= 1.76$$

$$p(P) \cdot p(y_1 = B, y_2 = 1, y_3 = 1|P) = \frac{5}{9} \cdot \frac{1}{5} \cdot 1.304$$

$$\approx 0.145$$

For class N:

$$p(y_3 = 1|N) = \frac{1}{0.1414\sqrt{2\pi}} \exp\left(-\frac{(1-1.0)^2}{2 \cdot 0.1414^2}\right)$$

$$= 2.82$$

$$p(N) \cdot p(y_1 = B, y_2 = 1, y_3 = 1|N) = \frac{4}{9} \cdot \frac{1}{4} \cdot 2.82$$

$$\approx 0.313$$

Since $0.145 < 0.313$**, the predicted class is N.**

(c) For observation (B, 0, 0.9):

For class P:

$$p(y_3 = 0.9|P) = \frac{1}{0.217\sqrt{2\pi}} \exp\left(-\frac{(0.9-0.82)^2}{2 \cdot 0.217^2}\right)$$

$$= 1.72$$

$$p(y_3 = 0.9) = \frac{1}{0.2\sqrt{2\pi}} \exp\left(-\frac{(0.9-0.9)^2}{2 \cdot 0.2^2}\right)$$

$$= 1.99$$

$$p(P) \cdot p(y_1 = B, y_2 = 0, y_3 = 0.9|P) = \frac{5}{9} \cdot \frac{1}{5} \cdot 1.72$$

$$\approx 0.191$$

For class N:

$$p(y_3 = 0.9|N) = \frac{1}{0.1414\sqrt{2\pi}} \exp\left(-\frac{(0.9-1.0)^2}{2 \cdot 0.1414^2}\right)$$

$$= 2.20$$

$$p(N) \cdot p(y_1 = B, y_2 = 0, y_3 = 0.9|N) = \frac{4}{9} \cdot \frac{2}{4} \cdot 2.20$$

$$\approx 0.489$$

Since $0.191 < 0.489$**, the predicted class is N.**

I.4 Solution:

	(A, 1, 0.8)	(B, 1, 1)	(B, 0, 0.9)
Class	P	N	N

5. Class-conditional frequency of each word in the training vocabulary.

c	"Amazing"	"run"	"I"	"like"	"it"	"Too"	"tired"	"bad"	N_c	V
P	1	1	1	1	1	0	0	0	5	8
N	0	1	0	0	0	1	1	1	4	

Under a ML assumption, for the word w:

$$\begin{aligned}\text{Predicted Class} &= \underset{c}{\text{argmax}}\{p(c|w)\} \\ &= \underset{c}{\text{argmax}}\{\frac{p(w|c)\cdot \cancel{p(c)}}{\cancel{p(w)}}\} \\ &= \underset{c}{\text{argmax}}\{\prod^{i} p(t_i|c)\}\end{aligned}$$

	"I"	"like"	"to"	"run"	$\prod^i p(t_i\|c)$
$freq(t_i\|P)$	1	1	0	1	$\frac{2^3\cdot 1}{13^4} = 0.000280$
$p(t_i\|P)$	$\frac{1+1}{5+8}$	$\frac{1+1}{5+8}$	$\frac{0+1}{5+8}$	$\frac{1+1}{5+8}$	
$freq(t_i\|N)$	0	0	0	1	$\frac{2\cdot 1^3}{12^4} = 0.000096$
$p(t_i\|N)$	$\frac{0+1}{4+8}$	$\frac{0+1}{4+8}$	$\frac{0+1}{4+8}$	$\frac{1+1}{4+8}$	

I.5 Solution:
Since $0.000280 > 0.000096$**, the predicted class is P.**

Part II: Programming

1. (a) Boxplot comparison of the KNN and the Naive Bayes with Gaussian Assumption:

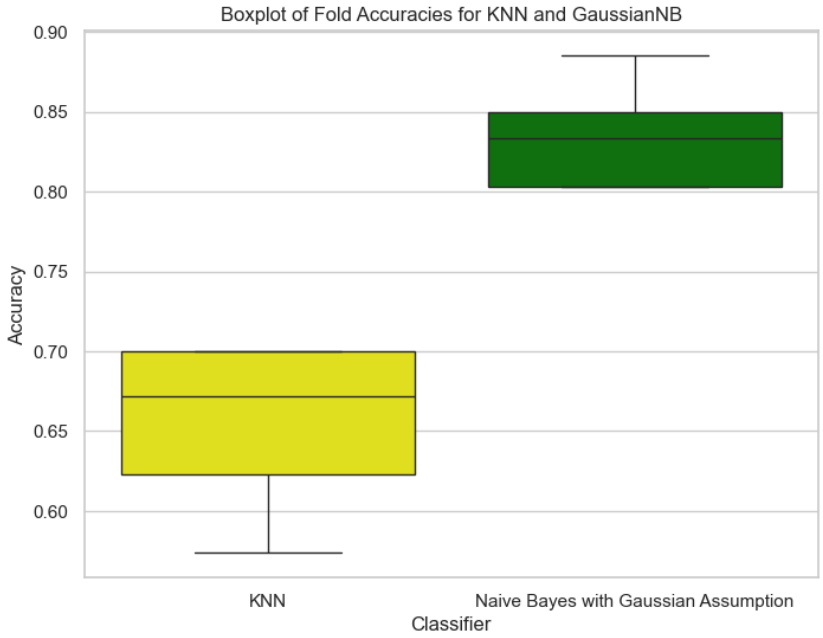


Figure 1: Boxplots

The KNN had an overall lower accuracy and was less stable regarding the performance results compared to the Naive Bayes model. This is likely because of the dataset was not scaled. Based on the fact that a knn makes prediction based on the "closest points" to an instance, if the features are on different scales, one feature with a large range will dominate the distance calculation, which can distort the actual "closeness" between points and therefore cause a unstability in the performance of the KNN.

(b) Original KNN score: 0.65 ± 0.05
KNN score on scaled dataset: 0.82 ± 0.02
Original Naive Bayes score: 0.84 ± 0.03
Naive Bayes score on scaled dataset: 0.84 ± 0.03

The KNN model had significant improvement in accuracy and had more stable results after min-max scaling the dataset. As mentioned in the previous answer the K-nearest neighbour models are sensitive to the scale of the dataset, and scaling can significantly improve the performance of the model. Scaling ensures that all features contribute equally to the distance calculation in the KNN.

The Naive bayes model with a gaussian assumption had very little to no difference in the results between the scaled and non-scaled data. In the Gaussian model, each feature is modeled separately with its own mean (μ) and variance (σ) according to the normal distribution. The Naive Bayes model with a Gaussian assumption

is insensitive to scaling because this mean and variance are put into consideration when calculating predictions. This explains the indifference in the results.

(c) $H_0 : p_1 = p_2$; $H_1 : p_1 > p_2$; pvalue= 0.7462688051215336
Through the statistical hypothesis test we computed a very high p-value (pvalue=0.7462688), and therefore cannot reject the null hypothesis at any of the common significance levels (0.1, 0.05, 0.01). This means that we do not have enough statistical evidence supporting the assertion that "the kNN model is statistically superior to naïve Bayes regarding accuracy".

2. (a) Train and test accuracies:

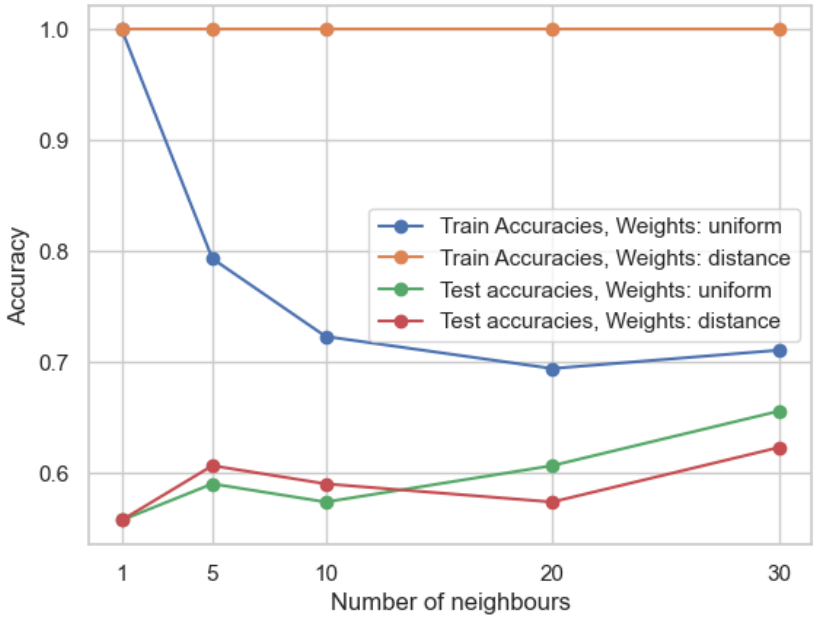
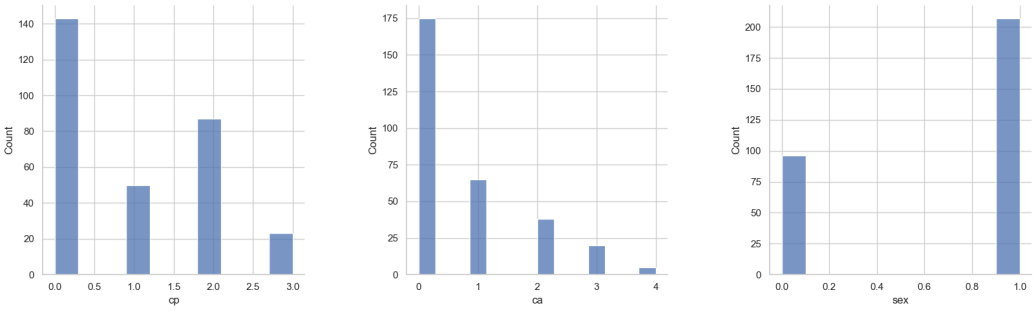


Figure 2: KNN accuracies

(b) Increasing the number of neighbours, in the knns of both weights (Uniform and Distance), improves the generalization capacity of the model. The model tends to overfit the smaller the number of neighbours parameter is. This is because when you increase n_neighbors, the model considers more neighboring data points to make a prediction. This therefore prevents overfitting because the decision is based on a larger sample of the training data, rather than on a few close neighbors.

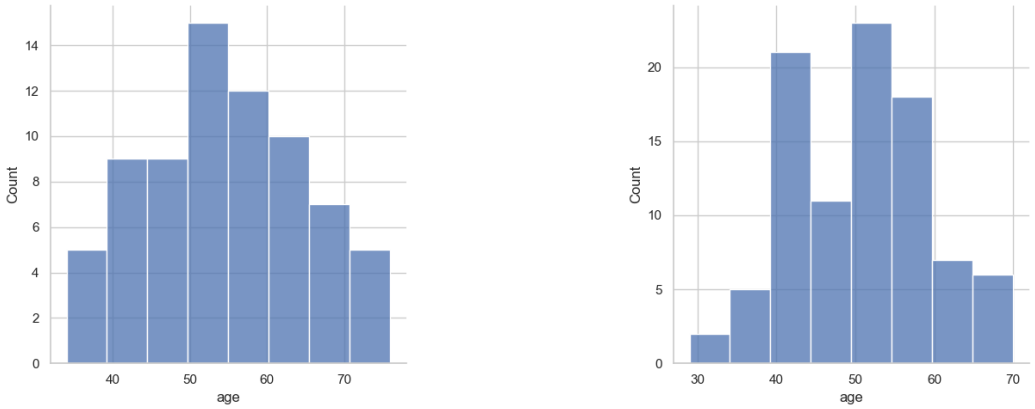
3. The Naive Bayes model assumes that the features of the data have a distribution similar to that of a normal distribution. However, some features do not conform to normal distributions well.



(a) cp feature distribution (b) ca feature distribution (c) sex feature distribution

Figure 3: Feature distributions

The model also assumes that the features are independent to one another. However, some features appear to not be independent to one another. For example those with heart disease who have sex == 1, seem to tend to have a younger age than those with sex == 0.



(a) age distribution with sex==0 (b) age distribution with sex==1

Figure 4: Feature distributions

After making a statistical test of such hypothesis and obtaining a pvalue of 0.007 we can say that evidence supports it being true even for common significance levels of $\alpha = 0.01$.