Aprendizagem 2023
Homework I — Group 003
(ist1107028, ist1107137)

Part I: Pen and paper

We collected four positive (P) observations,

$$x_1 = (A, 0), \quad x_2 = (B, 1), \quad x_3 = (A, 1), \quad x_4 = (A, 0)$$

and four negative (N) observations,

$$x_5 = (B, 0), \quad x_6 = (B, 0), \quad x_7 = (A, 1), \quad x_8 = (B, 1)$$

Consider the problem of classifying observations as positive or negative.

1. **Compute the F1-measure of a kNN with $k = 5$ and Hamming distance using a leave-one-out evaluation schema. Show all calculus.**

   We start by calculating Hamming distance between observations. The Hamming distance is the number of positions at which the corresponding symbols are different.
   Since we are workin with $k = 5$, we will consider the 5 nearest neighbors of each observation (written in blue).

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
x_1	-	2	**1**	**0**	**1**	**1**	**1**	2
x_2	2	-	**1**	2	**1**	**1**	**1**	**0**
x_3	**1**	**1**	-	**1**	2	2	**0**	**1**
x_4	**0**	2	**1**	-	**1**	**1**	**1**	2
x_5	**1**	**1**	2	**1**	-	**0**	2	**1**
x_6	**1**	**1**	2	**1**	**0**	-	2	**1**
x_7	**1**	**1**	**0**	**1**	2	2	-	**1**
x_8	2	**0**	**1**	2	**1**	**1**	**1**	-

Table 1: Hamming distance between observations

   Now that we have the Hamming distance between all observations, we must identify if the prediction is correct or not. We will consider the majority class of the 5 nearest neighbors for each observation.

   Example: For x_1, the 5 nearest neighbors are x_3 and x_4 (which are positive), x_5, x_6 and x_7 (which are negative). The majority class is negative, therefore the prediction is incorrect.

We apply the same logic for the rest of the classes, ending up with the following table:

Observation	True Value	Prediction	Confusion Matrix Terminology
x_1	P	N	FP
x_2	P	N	FP
x_3	P	P	TP
x_4	P	N	FN
x_5	N	P	FP
x_6	N	P	FP
x_7	N	P	FP
x_8	N	N	TN

P - Positive observation; N - Negative observation

TP - True Positive; TN - True Negative; FP - False Positive; FN - False Negative

Table 2: Predictions for each observation

With this table, we can know calculate the Precision, Recall and F1-measure using the following formulas:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{1}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{2}$$

$$\text{F1-measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

Replacing the corresponding values in the formulas, we get:

for Precision (1) and Recall (2):

$$\text{Precision} = \frac{1}{1+5} \approx 0.1667 \qquad \text{Recall} = \frac{1}{1+1} = 0.5$$

F1-measure (3):

$$\text{F1-measure} = 2 \times \frac{0.1667 \times 0.5}{0.1667 + 0.5} \approx 0.25 \tag{4}$$

2. **Propose a new metric (distance) that improves the latter's performance (i.e., the F1-measure) by three fold.**