

## Aprendizagem 2021/22

**Homework I**

Deadline 18/10/2021 23:59 via Fenix as PDF

- Homework limited to 4 pages (2.5–3pp for part I, 1–1.5pp for part II) according to the provided template
- Include your programming code as an Appendix (maximum 1 page)
- Submission Gxxx.PDF in Fenix where xxx is your group number. Please note that it is possible to submit several times on Fenix to prevent last-minute problems. Yet, only the last submission is considered valid
- Exchange of ideas is encouraged. Yet, if copy is detected after automatic/manual clearance, homework is nullified and IST guidelines apply for content sharers and consumers, irrespectively of the underlying intent
- Please consult the FAQ before posting questions to your faculty hosts

**I. Pen-and-paper [12.5v]**

Considering the following training data:

|          | $y_1$ | $y_2$ | $y_3$ | $y_4$ | <i>class</i> |
|----------|-------|-------|-------|-------|--------------|
| $x_1$    | 0.6   | A     | 0.2   | 0.4   | 0 (N)        |
| $x_2$    | 0.1   | B     | -0.1  | -0.4  | 0            |
| $x_3$    | 0.2   | A     | -0.1  | 0.2   | 0            |
| $x_4$    | 0.1   | C     | 0.8   | 0.8   | 0            |
| $x_5$    | 0.3   | B     | 0.1   | 0.3   | 1 (P)        |
| $x_6$    | -0.1  | C     | 0.2   | -0.2  | 1            |
| $x_7$    | -0.3  | C     | -0.1  | 0.2   | 1            |
| $x_8$    | 0.2   | B     | 0.5   | 0.6   | 1            |
| $x_9$    | 0.4   | A     | -0.4  | -0.7  | 1            |
| $x_{10}$ | -0.2  | C     | 0.4   | 0.3   | 1            |

- 1) [4.5v] Train a Bayesian classifier assuming: i) independence and equal importance between  $\{y_1\}$ ,  $\{y_2\}$  and  $\{y_3, y_4\}$  variable sets, and ii) numeric variable sets are normally distributed.
- 2) [4.5v] Draw a confusion matrix for the training observations.  
Note: you can use programming packages to support your calculus, yet show intermediary results.
- 3) [1v] Evaluate the training F1 score.
- 4) [2.5v] Identify the decision probability threshold that optimizes training accuracy. Comment.

**II. Programming and critical analysis [7.5v]**Considering the `breast.w.arff` dataset available at the Homeworks tab in the course webpage

- 5) [1.5v] Draw the class-conditional distributions per variable using a 3x3 plot grid.
- 6) [3v] Using a 10-fold cross validation with `seed=<group number>`, assess the accuracy of  $k$ NN under  $k \in \{3, 5, 7\}$ , Euclidean distance and uniform weights. Show empirically, which  $k$  is less susceptible to the overfitting risk?
- 7) [1.5v] Fixing  $k = 3$ , and assuming accuracy estimates are normally distributed, test the hypothesis “ $k$ NN is statistically superior to Naïve Bayes (multinomial assumption)”.
- 8) [1.5v] Given the empirical data collected along 5-7, enumerate two reasons that can underlie the differences in performance between  $k$ NN and Naïve Bayes.

**END**