TÉCNICO LISBOA

Aprendizagem 2023/24

# Homework II

Deadline 9/10/2023 (Monday) 23:59 via Fenix as PDF

- Submit Gxxx.ZIP in Fenix where xxx is your group number. The ZIP should contain two files: Gxxx_report.pdf with your report and Gxxx_notebook.ipynb with your notebook demo according to the suggested template
- It is possible to submit several times on Fenix to prevent last-minute problems. Yet, only the last submission is kept
- Exchange of ideas is encouraged. Yet, if copy is detected after automatic or 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 [13v]

Consider the following dataset:

| $D$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ | $y_6$ |
|---|---|---|---|---|---|---|
| $\mathbf{x}_1$ | 0.24 | 0.36 | 1 | 1 | 0 | A |
| $\mathbf{x}_2$ | 0.16 | 0.48 | 1 | 0 | 1 | A |
| $\mathbf{x}_3$ | 0.32 | 0.72 | 0 | 1 | 2 | A |
| $\mathbf{x}_4$ | 0.54 | 0.11 | 0 | 0 | 1 | B |
| $\mathbf{x}_5$ | 0.66 | 0.39 | 0 | 0 | 0 | B |
| $\mathbf{x}_6$ | 0.76 | 0.28 | 1 | 0 | 2 | B |
| $\mathbf{x}_7$ | 0.41 | 0.53 | 0 | 1 | 1 | B |
| $\mathbf{x}_8$ | 0.38 | 0.52 | 0 | 1 | 0 | A |
| $\mathbf{x}_9$ | 0.42 | 0.59 | 0 | 1 | 1 | B |

1. Consider $\mathbf{x}_1$–$\mathbf{x}_7$ to be training observations, $\mathbf{x}_8$–$\mathbf{x}_9$ to be testing observations, $y_1 - y_5$ to be input variables and $y_6$ to be the target variable.
   *Hint*: you can use `scipy.stats.multivariate_normal` for multivariate distribution calculus

   a. [3.5v] Learn a Bayesian classifier assuming: i) $\{y_1, y_2\}$, $\{y_3, y_4\}$ and $\{y_5\}$ sets of independent variables (e.g., $y_1 \perp\!\!\!\perp y_3$ yet $y_1 \perp\!\!\!\perp y_2$), and ii) $y_1 \times y_2 \in \mathbb{R}^2$ is normally distributed. Show all parameters (distributions and priors for subsequent testing).

   b. [2.5v] Under a MAP assumption, classify each testing observation showing all your calculus.

   c. [2v] Consider that the default decision threshold of $\theta = 0.5$ can be adjusted according to

   $$f(\mathbf{x}|\theta) = \begin{cases} \text{A} & P(\text{A}|\mathbf{x}) > \theta \\ \text{B} & \text{otherwise} \end{cases}.$$

   Under a maximum likelihood assumption, what thresholds optimize testing accuracy?

2. Let $y_1$ be the target numeric variable, $y_2$-$y_6$ be the input variables where $y_2$ is binarized under an equal-width (equal-range) discretization. For the evaluation of regressors, consider a 3-fold cross-validation over the full dataset ($\mathbf{x}_1$- $\mathbf{x}_9$) without shuffling the observations.

   a. [1v] Identify the observations and features per data fold after the binarization procedure.

   b. [4v] Consider a distance-weighted $k$NN with $k = 3$, Hamming distance ($d$), and $1/d$ weighting. Compute the MAE of this $k$NN regressor for the 1st iteration of the cross-validation (i.e. train observations have the lower indices).

Aprendizagem 2023/24

**Homework II**

Deadline 9/10/2023 (Monday) 23:59 via Fenix as PDF

## II. Programming and critical analysis [7v]

Considering the `column_diagnosis.arff` dataset available at the course webpage's homework tab. Using `sklearn`, apply a 10-fold stratified cross-validation with shuffling (`random_state=0`) for the assessment of predictive models along this section.

1) [3v] Compare the performance of $k$NN with $k = 5$ and naïve Bayes with Gaussian assumption (consider all remaining parameters for each classifier as `sklearn`'s default):

    a. Plot two boxplots with the fold accuracies for each classifier.

    b. Using `scipy`, test the hypothesis "$k$NN is statistically superior to naïve Bayes regarding accuracy", asserting whether is true.

2) [2.5v] Consider two $k$NN predictors with $k = 1$ and $k = 5$ (uniform weights, Euclidean distance, all remaining parameters as default). Plot the differences between the two cumulative confusion matrices of the predictors. Comment.

3) [1.5v] Considering the unique properties of `column_diagnosis`, identify three possible difficulties of naïve Bayes when learning from the given dataset.

**END**