TÉCNICO LISBOA

Aprendizagem 2022/23

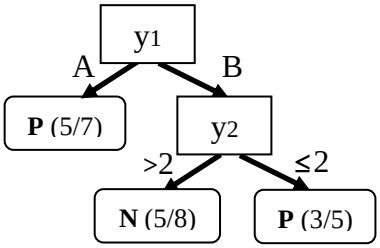
Homework I

Deadline: 7/10/2022 (Friday) 23:59 via Fenix as PDF

- 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
- Use the provided report template. Include your programming code as an Appendix
- 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 [12v]

Given the following decision tree learnt from 20 observation using Shannon entropy, with leaf annotations (`#correct/#total`)



1) [4v] Draw the training confusion matrix.
2) [3v] Identify the training F1 after a post-pruning of the given tree under a maximum depth of 1.
3) [2v] Identify two different reasons as to why the left tree path was not further decomposed.
4) [3v] Compute the information gain of variable y1.

II. Programming [8v]

Considering the `pd_speech.arff` dataset available at the homework tab:

1) [6v] Using `sklearn`, apply a stratified 70-30 training-testing split with a fixed seed (`random_state=1`), and assess in a single plot the training and testing accuracies of a decision tree with no depth limits (and remaining default behavior) for a varying number of selected features in {5,10,40,100,250,700}. Feature selection should be performed before decision tree learning considering the discriminative power of the input variables according to mutual information criterion (`mutual_info_classif`).
2) [2v] Why training accuracy is persistently 1? Critically analyze the gathered results.

END