TÉCNICO LISBOA

Aprendizagem 2025/26

Homework I

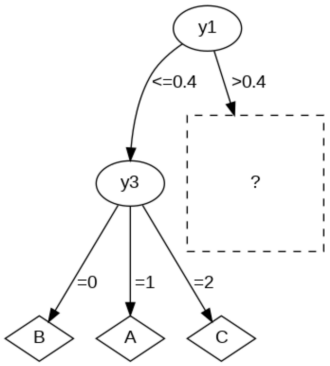
Deadline: via Fenix as PDF

- 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 templates
- 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 [11v]

Consider the partially learnt decision tree from the dataset D. D is described by 4 input variables (1 numeric and 3 categorical – and a target variable with three classes).

D	y_1	y_2	y_3	y_4	y_{out}
x_1	0.12	1	1	0	A
x_2	0.18	0	0	1	B
x_3	0.25	2	2	1	C
x_4	0.33	0	1	0	A
x_5	0.45	1	0	2	C
x_6	0.52	0	0	0	B
x_7	0.58	2	1	2	C
x_8	0.62	1	0	1	A
x_9	0.71	1	2	1	A
x_{10}	0.83	1	2	1	B
x_{11}	0.90	2	1	2	B
x_{12}	0.95	2	2	2	C



1) [5v] Complete the given decision tree using information gain (Shannon entropy with log_2) and considering that i) a minimum of 4 observations is required to split an internal node, and ii) decisions by ascending alphabetic should be placed in case of ties.

2) [2.5v] Draw the training confusion matrix for the learnt decision tree.

3) [1.5v] Identify which class has the lowest training F1 score.

4) [1v] consider 2 new observations: x_{13}= [0.02, 1, 2, 1, A] and x_{14} = [0.27, 2, 0, 0, C]. Draw the class-conditional relative histograms of y_1 using 5 equally spaced bins in [0,1]. Find the n-ary root split using the discriminant rules from these empirical distributions.

5) [1v] Identify the outliers in this dataset.

TÉCNICO LISBOA

Aprendizagem 2025/26

Homework I

Deadline: via Fenix as PDF

II. Programming [9v]

Consider the hungarian_heart_diseases.csv data available at the homework tab, comprising 9 biological features to classify 284 patients into 2 classes (normal, heart disease).

1) [3v] Using a stratified 80-20 training-testing split with a fixed seed (random_state=1), assess in a single plot both the training and testing accuracies of a decision tree with minimum sample leaf in {1, 3, 5, 10, 25, 50, 100} and the remaining parameters as default.

2) [2v] Critically analyze these results, including the generalization capacity across settings.

3) [4v] A healthcare provider requested the development of a predictive model achieving at least *80% validation accuracy and 78.5% test accuracy*. The dataset must be split into training, validation, and testing sets (60–20–20), using a stratified split. The goal is to identify a model that satisfies the required accuracy within the following hyperparameter ranges: max_depth $\in$ [2,4] and min_samples_split $\in$ [2,100]. All other hyperparameters should be kept at their default values. Use random_state = 1 to split the data and to create the model.

    i. Plot the decision tree.

    ii. Explain what characterizes heart diseases by identifying the conditional associations together with their posterior probabilities.

END