Aprendizagem 2024/25

Homework I

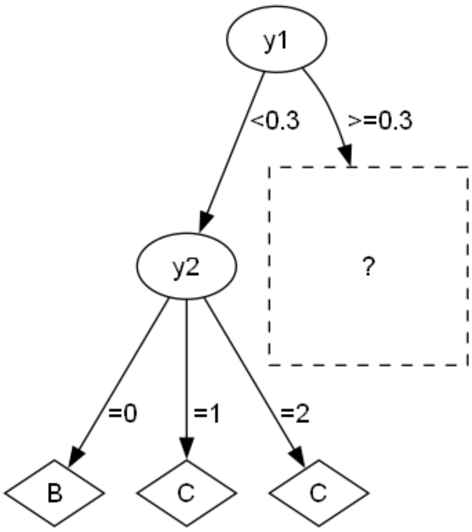
Deadline: 27/9/2024 (Friday) 23:59 via Fenix as PDF

- 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 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 four input variables – one numeric with values in [0,1] and 3 categorical – and a target variable with three classes.

D	y_1	y_2	y_3	y_4	y_{out}
x_1	0.22	2	0	1	C
x_2	0.06	0	0	0	B
x_3	0.16	1	2	2	C
x_4	0.21	0	0	0	B
x_5	0.01	2	2	0	C
x_6	0.3	0	1	0	B
x_7	0.76	0	1	1	A
x_8	0.86	1	0	0	A
x_9	0.93	0	1	1	C
x_{10}	0.47	0	1	1	C
x_{11}	0.73	1	0	0	A
x_{12}	0.89	1	2	0	B



TÉCNICO LISBOA

Aprendizagem 2024/25

Homework I

Deadline: 27/9/2024 (Friday) 23:59 via Fenix as PDF

1) [5v] Complete the given decision tree using Shannon entropy (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) [2v] Draw the class-conditional relative histograms of y1 using 5 equally spaced bins in [0,1]. Find the n-ary root split using the discriminant rules from these empirical distributions.

II. Programming [9v]

Consider the `diabetes.arff` data available at the homework tab, comprising 8 biological features to classify 768 patients into 2 classes (normal, diabetes).

1) [1v] ANOVA is a statistical test that can be used to assess the discriminative power of a single input variable. Using `f_classif` from `sklearn`, identify the input variables with the worst and best discriminative power. Plot their class-conditional probability density functions.

2) [4v] 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 split in $\{2, 5, 10, 20, 30, 50, 100\}$ and the remaining parameters as default.

   [optional] Note that split thresholding of numeric variables in decision trees is non-deterministic in sklearn, hence you may opt to average the results using 10 runs per parameterization.

3) [2v] Critically analyze these results, including the generalization capacity across settings.

4) [2v] To deploy the predictor, a healthcare provider opted to learn a single decision tree (`random_state=1`) using *all* available data and ensuring that the maximum depth would be 3 in order to avoid overfitting risks.

   i. Plot the decision tree.

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

END