TÉCNICO LISBOA

Aprendizagem 2023/24

Homework I

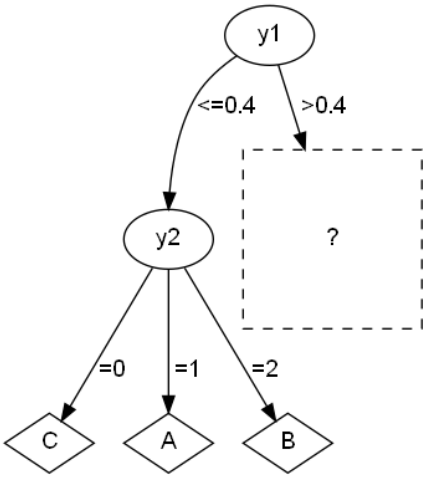
Deadline: 29/9/2023 (Friday) 23:59 via Fenix as PDF

- Submit Gxxx.ZIP in Fenix where xxx is your group number. The ZIP should contain two files: Gxxx.pdf with your report and Gxxx.ipynb with your python notebook
- 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}
$\mathbf{x}_1$	0.24	1	1	0	A
$\mathbf{x}_2$	0.06	2	0	0	B
$\mathbf{x}_3$	0.04	0	0	0	B
$\mathbf{x}_4$	0.36	0	2	1	C
$\mathbf{x}_5$	0.32	0	0	2	C
$\mathbf{x}_6$	0.68	2	2	1	A
$\mathbf{x}_7$	0.9	0	1	2	A
$\mathbf{x}_8$	0.76	2	2	0	A
$\mathbf{x}_9$	0.46	1	1	1	B
$\mathbf{x}_{10}$	0.62	0	0	1	B
$\mathbf{x}_{11}$	0.44	1	2	2	C
$\mathbf{x}_{12}$	0.52	0	2	0	C



1) [5v] Complete the given decision tree using Information gain with Shannon entropy ($\log_2$). Consider that: i) a minimum of 4 observations is required to split an internal node, and ii) decisions by ascending alphabetic order 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] Considering y_2 to be ordinal, assess if y_1 and y_2 are correlated using the Spearman coefficient.

5) [1v] Draw the class-conditional relative histograms of y_1 using 5 equally spaced bins in [0,1]. Challenge: find the root split using the discriminant rules from these empirical distributions.

TÉCNICO LISBOA

Aprendizagem 2023/24

Homework I

Deadline: 29/9/2023 (Friday) 23:59 via Fenix as PDF

II. Programming [9v]

To answer the following questions, consider using the `sklearn` API documentation and the notebooks in the course webpage as guidance. Show in your PDF report both the code and the corresponding results.

Consider the `column_diagnosis.arff` data available at the homework tab, comprising 6 biomechanical features to classify 310 orthopaedic patients into 3 classes (`normal, disk hernia, spondilolysthesis`).

1) [1.5v] Apply `f_classif` from `sklearn` to assess the discriminative power of the input variables. Identify the input variable with the highest and lowest discriminative power.
   Plot the class-conditional probability density functions of these two input variables.

2) [4v] Using a stratified 70-30 training-testing split with a fixed seed (`random_state=0`), assess in a single plot both the training and testing accuracies of a decision tree with depth limits in {1,2,3,4,5,6,8,10} 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) [1.5v] Comment on the results, including the generalization capacity across settings.

4) [2v] To deploy the predictor, a healthcare team opted to learn a single decision tree (`random_state=0`) using *all* available data as training data, and further ensuring that each leaf has a minimum of 20 individuals in order to avoid overfitting risks.

   i. Plot the decision tree.

   ii. Characterize a hernia condition by identifying the hernia-conditional associations.

END