Aprendizagem 2024

Lab 2: Decision Trees

Practical exercises

1. Consider the following dataset:

	y_1	y_2	y_3	class
x_1	a	a	a	+
x_2	c	b	c	+
x_3	c	a	c	+
x_4	b	a	a	−
x_5	a	b	c	−
x_6	b	b	c	−

Plot the learned decision tree using information gain (Shannon entropy). Show your calculus.

2. Show if a decision tree can learn the following logical functions and, if so, plot the corresponding decision boundaries.
    a) AND
    b) OR
    c) XOR

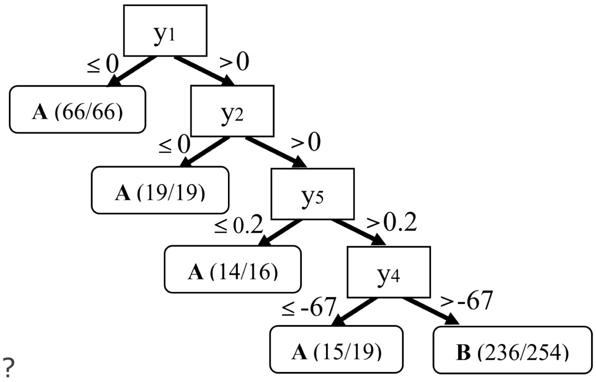
3. Consider the following testing targets, z, and the corresponding predictions, $\hat{z}$, by a decision tree:

$$z = [AAABBBCCCC]$$
$$\hat{z} = [BBACBACABC]$$

    a) Draw the confusion matrix
    b) Compute the accuracy and sensitivity per class
    c) Considering class C, identify its precision and F_1-measure
    d) Identify the accuracy, sensitivity, and precision of a random classifier

4. Consider a dataset composed by 374 records, described by 6 variables, and classified according to the following

decision tree. Each leaf in the tree shows the label, number of classified records with the label, and total number of observations in the leaf. The positive class is the minority class.



a) Compute the confusion matrix.

b) Compare the accuracy of the given tree versus a pruned tree with only two nodes.

   Is there any evidence towards overfitting?

c) [*optional*] Are decision trees learned from high-dimensional data susceptible to underfitting?

   Why an ensemble of decision trees minimizes this problem?

Programming quests

5. Following the provided Jupyter notebook on *Classification*, learn and evaluate a decision tree classifier on the *breast.w.arff* dataset (available at the webpage) using *sklearn*.

   Considering a 80-20 train-test split:

   a) visualize the decision tree learned from the training observations with default parameters

   b) compare the train and test accuracy of decision trees with a maximum depth in {1, 2, 3}