Aprendizagem 2023

Lab 2: Decision Trees

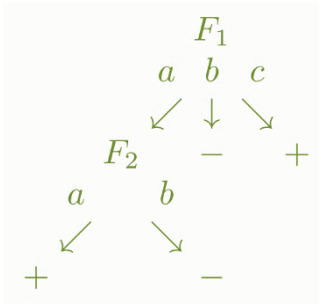
Practical exercises

I. Decision tree learning

1. Consider the following dataset:

	y_1	y_2	y_3	class
$\mathbf{x}_1$	a	a	a	+
$\mathbf{x}_2$	c	b	c	+
$\mathbf{x}_3$	c	a	c	+
$\mathbf{x}_4$	b	a	a	−
$\mathbf{x}_5$	a	b	c	−
$\mathbf{x}_6$	b	b	c	−

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



Brief notes:

y_1 provides the highest gain, $IG(y_{out}|y_1) = 1 - 0.33$, hence selected.

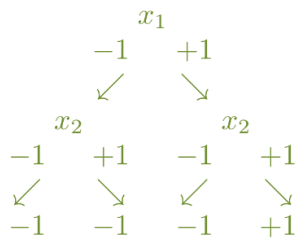
y_1 correctly classifies all observations when $y_1 = b$ and $y_1 = c$

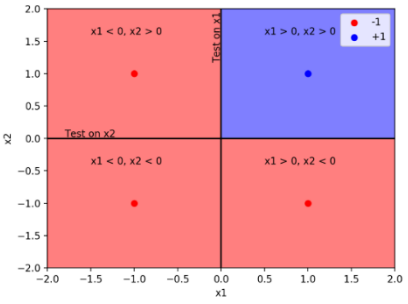
Entropies of y_2 and y_3 for $(y_1 = a)$-conditional data are both zero, so we can select either.

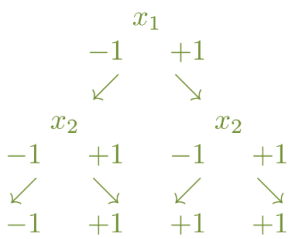
There is no more uncertainty.

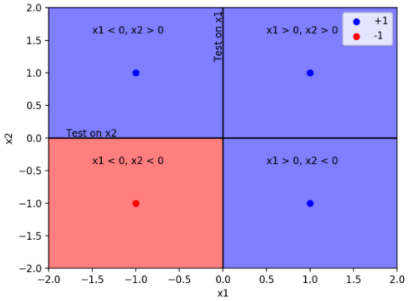
2. Show if a decision tree can learn the following logical functions and, if so, plot the corresponding decision boundaries.

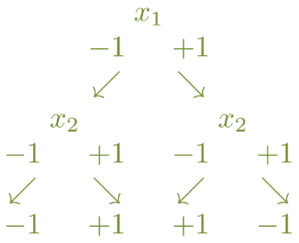
a) AND
b) OR
c) XOR

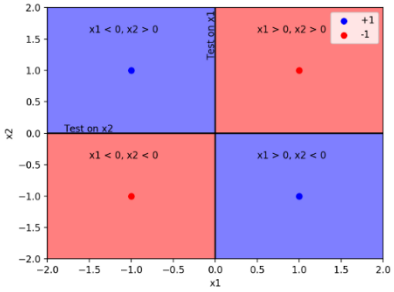












3. Consider the following testing targets, z, and the corresponding predictions, $\hat{z}$, by a decision tree:

$$z = [A\ A\ A\ B\ B\ B\ C\ C\ C\ C]$$
$$\hat{z} = [B\ B\ A\ C\ B\ A\ C\ A\ B\ C]$$

a) Draw the confusion matrix

		true		
		A	**B**	**C**
predicted	**A**	1	1	1
	B	2	1	1
	C	0	1	2

b) Compute the accuracy and sensitivity/recall per class

$$acccuracy = 0.4, sensitivity_A = \frac{1}{3}, sensitivity_B = \frac{1}{3}, sensitivity_C = \frac{1}{2}$$

c) Considering class *C*, identify precision and F_1-measure

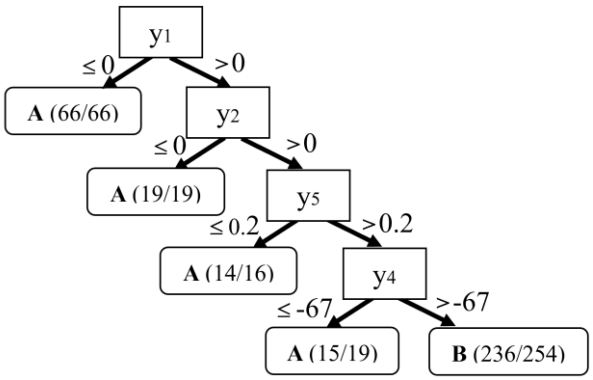
$$precision_C = \frac{2}{3}, \qquad F1_C = 0.57$$

d) Identify the accuracy, sensitivity, and precision of the random classifier

$$accuracy_{random} = 0.(3), recall_{random}(A) = 0.(3), recall_{random}(B) = 0.(3), recall_{random}(C) = 0.(3)$$

$$precision_{random}(A) = 0.3, precision_{random}(B) = 0.3, precision_{random}(C) = 0.4$$

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

$$\#A = 66 + 19 + 14 + 15 + 18 = 132$$
$$\#B = 0 + 0 + 2 + 4 + 236 = 242$$

The minority class is A, hence is seen as positive.

		True	
		P (A)	**N** (B)
Predicted	**P** (A)	114	6
	N (B)	18	236

b) Compare the accuracy of the given tree versus a pruned tree with only two nodes.
Is there any evidence towards overfitting?

Considering training accuracy: $accuracy_{\text{depth=4}} = 0.936,\ \ accuracy_{\text{depth=2}} = 0.874$

Without the testing accuracy, there is no sufficient evidence to assume the tree is prone to overfit input data.

c) [*optional*] Are decision trees learned from high-dimensional data susceptible to underfitting?
Why an ensemble of decision trees minimizes this problem?

Assuming a limited depth, relevant data may be discarded due to a focus on a compact subset of overall input variables. In ensemble models, such as random forests, different decision trees can be learned from data subsamples and subspaces, leading to decisions that consider a broader set of input variables.

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 varying maximum depth