Aprendizagem 2021

Lab 3: Information Theory and Decision Trees

Practical exercises

I. Data preprocessing and information theory

Consider the following dataset:

	y_1	y_2	y_3	y_4	y_{out}
x_1	0.2	0.5	A	A	A
x_2	0.1	-0.4	A	A	A
x_3	0.2	0.6	A	B	C
x_4	0.9	0.8	B	B	C
x_5	-0.3	0.3	B	B	B
x_6	-0.1	-0.2	B	B	B

where y_1 and y_2 are numeric variables in [-1,1], y_3 and y_4 are nominal, and y_{out} is ordinal

1. On feature importance and selection

a) According to Spearman, which numeric variable should be removed?

b) According to information gain, which nominal variable should be removed?

c) Assuming unsupervised feature importance, which numeric variable should be removed?

d) Using cross-entropy, assess the similarity between nominal variables.

2. Considering the following class estimates for two observations x_1 and x_2:

$$p(A|\mathbf{x}_1) = 0.1,\ p(B|\mathbf{x}_1) = 0.6,\ p(A|\mathbf{x}_2) = 0.3,\ p(B|\mathbf{x}_2) = 0.2.$$

Considering cross-entropy, which one of the observations is more adequately classified?

3. Normalize y_2 using min-max scaling and standardization. Compare the results

4. Binarize y_1 considering

a) y_1 follows a Uniform distribution (equal-width/range discretization)

b) a balanced distribution of bins (equal-depth/frequency discretization)

II. Decision tree learning

5. Consider the following data table:

	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	−

a) Plot the learned decision tree using ID3 with information gain.

Show the calculus for the first split and the intuition for the remaining steps.

b) Repeat the same exercise but this time using CART with the Gini index.

c) How do ID3, C4.5 and CART handle numeric variables?

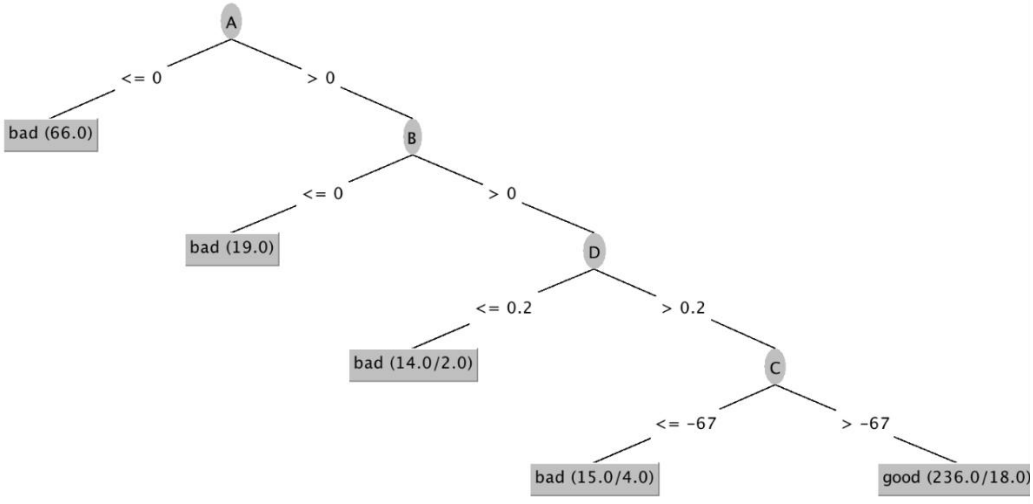
6. Show if a decision tree can learn the following logical functions and, if so, plot the corresponding decision boundaries.

a) AND

b) OR

c) XOR

7. Consider a dataset composed by 350 records, described by 6 variables, and classified according to the decision tree in **Error! Reference source not found.**. Each leaf in the tree shows the label, number of classified records with the label, and the number of errors covered 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) Are decision trees learned from high-dimensional data susceptible to underfitting?

Why an ensemble of decision trees minimizes this problem?

Programming quests

Reuse the *sklearn* code from last practical classes on the iris data classification.

8. 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}
   c) statistically test accuracy differences between decision trees and random forests using CV

9. Rank iris data features according to mutual information (*mutual_info_classif*)
Resource: https://scikit-learn.org/stable/modules/feature_selection.html