Aprendizagem - First Assignment

Ruben Teimas, m47753

June 2021

1 Introduction

In this first assignment of Aprendizagem we were asked to implement a decision tree using the *ID3* algorithm. Additionally this decision tree should support at least 2 purity functions and one form of post-pruning and another form of pre-pruning.

2 Development

The first step, before implementing the tree, was to read the support material from classes in order to better understand the process.

After that I thought about the purity functions to use. The chosen ones were the *information gain*, which was the original function used in the *ID3* algorithm, and *gini index*. *Gini* is an improvement over *informtion gain* because it computationally less expensive, since it does not use logarithmic functions.

For the pruning, I thought about using *Max depth limit* (pre-pruning) and *Reduced Error* (post-pruning).

3 Implementation

In order to represent a decision tree compatible with *sklearn* environment, I created the following classes:

- DecisionTree.
- Node.

The *DecisionTree* class has, as attributes, the *root node*, the purity function, the pruning type, the depth for the current branch, and the number of hits (correctly classified entries).

The tree's *root* is an instance of the *Node* class. This class has the class value, which is the class value for the leafs and the arc that led to it, for the remaining nodes. It has the attribute chosen for that node. Finally, it has a dictionary of children.

This dictionary has the arc (class value) as the key and the node reference as the value.

4 Results

The tables bellow show the score for both measures. For each measure the tree was fitted and scored 5 times, making the mean of the results.

Dataset	Information Gain	Gini index
weather.nominal	0,80	0,90
soybean	0,67	0,73
contact-lenses	0,69	0,73
vote	0,95	0,96

Table 1: Accuracy wihtout any pruning.

Dataset	Information Gain	Gini index
weather.nominal	0,80	0,90
soybean	0,47	0,14
contact-lenses	0,59	0,59
vote	0,90	0,91

Table 2: Accuracy using max depth limit=5.

By looking at the results it is possible to see that, when pre-pruning the tree, the results were overall worst. However, it is noticeable that the tree processing is faster than without the pruning.

Even with above statement in consideration, the soybean accuracy when using *gini* and *pre-pruning* is odd and it indicates that the tree does not generalize well for that dataset with a maximum depth of 5.

5 Considerations

The assignment was partially completed with success. The implementation gave me a greater notion of how decision trees work because I had to think about it in order to make the best implementation decisions.

Unfortunately I was not able to implement the post-pruning feature that I initially intended due to lack of time.