Norwegian University of Science and Technology

TDT4171 - Assignment 6

Artificial Intelligence Methods

Implementing a decision tree

Group members
Zachari Thiry

Group number: No Group

March 1, 2023 (Spring 2020)

Comments: Code is given in the link below.

https://github.com/Zachari-THIRY/TDT-4171/tree/main/Assignment6

NTNU
Norwegian University of
Science and Technology

Introduction

In this assignment, we are tasked to implement a decision tree based on a given set of helper functions.

1 Exercise 1

The given code has been fully implemented, with two possible importance functions : "random" and "information_gain". Below, I describe the behaviour of the two :

1.1 "Information Gain"

The "information_gain" option of importance() has been implemented according to Section 19 of the recommended book. This option of importance is deterministic in the sense that multiple runs will generate the same exact output. Before running, one could expect a fairly good score on the training accuracy, even a perfect score if the attributes can entirely dissociate the labels. Testing accuracy should be relatively good, or at least better than a random guess (0.75) to show that the tree is working.
Here are the results :

- training_accuracy = 1
- testing_accuracy = 0.923

Multiple runs of the algorithm confirm the deterministic character of "information_gain"

1.2 "random"

On this one, the output is randomized (as suggested by the name). For the training accuracy, one could expect it to still be equal to 1 since the tree has unlimited depth, and since separability is achievable (according to the information_gain results). Chances are that the testing accuracy will range anywhere between 0.75 (random guess) and 1.
Here are the results :

- training_accuracy : [1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0]
- testing_accuracy : [1.0, 1.0, 0.785, 0.928, 0.928, 0.928, 1.0, 0.928, 0.857, 0.857]

Our results coincide with our expectations : although the training accuracy remains at 1.0, meaning that there is good separability in the data set even for randomized choices, we notice fluctuation in the testing set.

1.3 Conclusions

Even though the training between "random" and "information_gain" are as efficient, there is no match for the "information_gain" when it comes to generalization capabilities.
In this assignment, the tree has no limit to it's depth (or 7 since we only have 7 attributes), but one could expect the information based tree to perform even better in cases where the tree size is constrained, or where one has to implement some nodes pruning.