TDT4171 Assignment 4

Anders Fagerli

March 20, 2019

1 Decision Tree Learning

The code appended implements a decision tree learning algorithm, with importance of each attribute calculated as expected information gain or a random attribute.

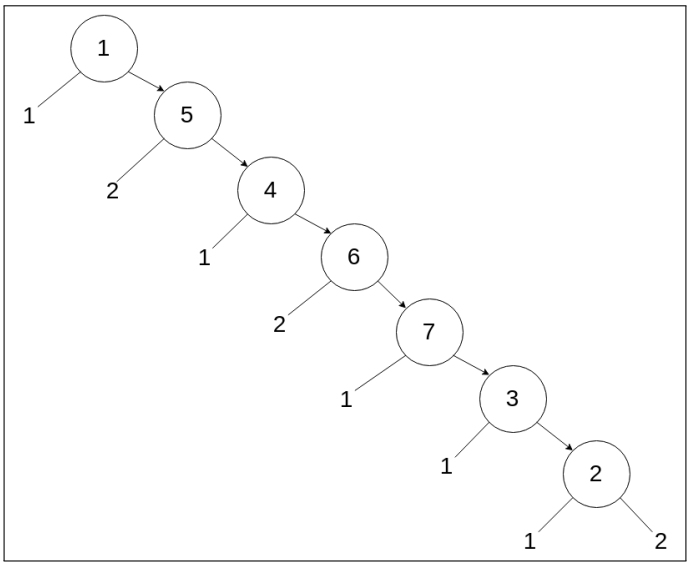

Figure 1: Decision tree with importance as expected information gain

From the tree above we can see that we have no reduction in attributes, meaning the tree is most likely wrong. This is a result of a bug in the code, which I have not found out. I will therefore try to discuss independently from the generated tree.

Using expected information gain, we expect to reduce the number of attributes needed to classify by choosing the most important attributes for the training set. This technique is deterministic, and will therefore produce the same classification rate every time we run it on a set of test data. As the most important attributes are chosen, this classifier performs well on the test set. It will also result in a smaller tree, reducing computational time for large sets.

A random choice of the attributes will produce a different result each time, and will on average perform worse than expected information gain. It will also result in larger trees, and therefore increase computational time.

Using the appended code, the algorithm with expected information gain gives a correct classification rate of 0.607, and the random number as importance gives a correct classification rate of around 0.75. This reinforces the indication of a fault in the code.