Naomi Igbinovia
Joshua Phillips
CSCI 4350
15 November 2023

Supervised Learning and the ID3 Decision Tree

Introduction

In machine learning, supervised learning is an incredibly useful approach in aiding machines in making predictions based on labeled data. Typically, a learning agent is given a labeled dataset to learn how to map out the data's features to its labels. The approach is based around the algorithm that will learn the relationship of the input features and the output answer to eventually predict new answers for brand-new inputs. One of the main techniques in supervised learning is the decision tree, which attempts to mimic human decision making. The decision tree is a type of supervised learning where the input data is recursively split into subsections based on the input features' values. As its name implies, the decision tree is structured as a tree-like graph where each node, aside from the leaves, represents a feature, each branch represents a decision based on said feature, and each leaf represents the classification. The goal of a decision tree is to build and correctly sort decisions to lead to the correct classification. In solving this algorithm, the id3.py program was developed.

The ID3 Method

Based around Iterative Dichotomiser 3 (ID3), this program takes a set of training data in the form of a text file, to build a decision tree based off of it. Then, a set of validation data also in the form of an input text file, is used to classify brand new data and test the performance of the built decision tree from the training the data. Using the ID3 algorithm, the program uses information gain to select the best attribute and split the data, carrying on the tree's structure until the full training set data is read in.

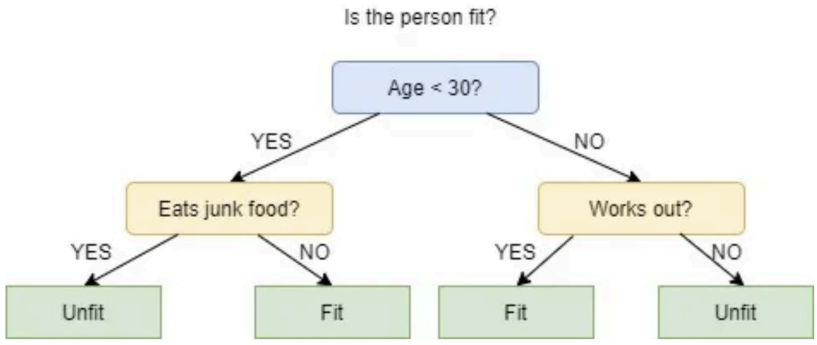

Figure 1. An ID3 decision tree.

Code Development

The id3.py program utilizes a class and a couple of methods prior to the main function. The DecisionNode class represents the given node in the decision tree. Each node of this class contains a Boolean checking if it's a leaf of not called 'terminal', a class label if the node is a leaf called 'category', a splitting value for non-leaf nodes called 'value', a feature index for splitting called 'feature', and references to left and right child nodes called 'left' and 'right'. The 'calculate_entropy(target_column)' method takes in the dataset's target column and returns its

entropy value using the Shannon entropy formula. The ‘calculate_info_gain(data, split_value, split_feature, target_feature)’ method takes in the given dataset, the split value, the split feature, and target feature, calculates the information gain for a potential split, and returns the information gain value. The ‘find_best_split(data, features, target)’ takes in the given data set, a features list, and the target feature index, finds the best split for the tree, and returns the best split. The 'find_split_with_max_gain(data, feature, target)’ method takes in the given data set, the feature index, and the target feature index, finds the split value that maximizes the information gain for a feature, and returns the maximum information gain and the matching split value. The 'build_decision_tree(data, features, target)’ method takes in the given dataset, a features list, and the target feature index, builds a decision tree recursively, and returns the tree’s root node. The 'test_decision_tree(test_data, target, tree)’ method takes in the test dataset, the target feature, and the decision tree, tests the decision on the test dataset, and returns the correct number of predictions.

$$H(x) = P(x)I(x) = -P(x) \cdot log_2(P(x))$$

Figure 2a. The Shannon Entropy formula.

$$I_{gain} = H(parent) - \sum_{i=1}^{n} [probability(S_i) \cdot H(S_i)]$$

$H(parent)$ – entropy of the parent node before the split.

$probability(S_i)$ – the probability of the ith child node after the split.

$H(S_i)$ – entropy of the ith child node after the split.

n – the number of child nodes after the split

Figure 2b. The Information Gain formula.

Performance

To test the performance of the ID3 algorithm, cross validation was used. The program ran in two sets of 100 rounds of training and test splits following the splits shown below. In this case, n represents the test set size

v = [1, 5, 10, 25, 50, 75, 100, 125, 140, 145, 149]

Figure 3a. The iris data set.

v = [1, 5, 10, 25, 50, 75, 90, 100, 104]

Figure 3b. The cancer data set.

The mean and standard error of the correct classification percentages were gathered for both datasets as charts displayed below.

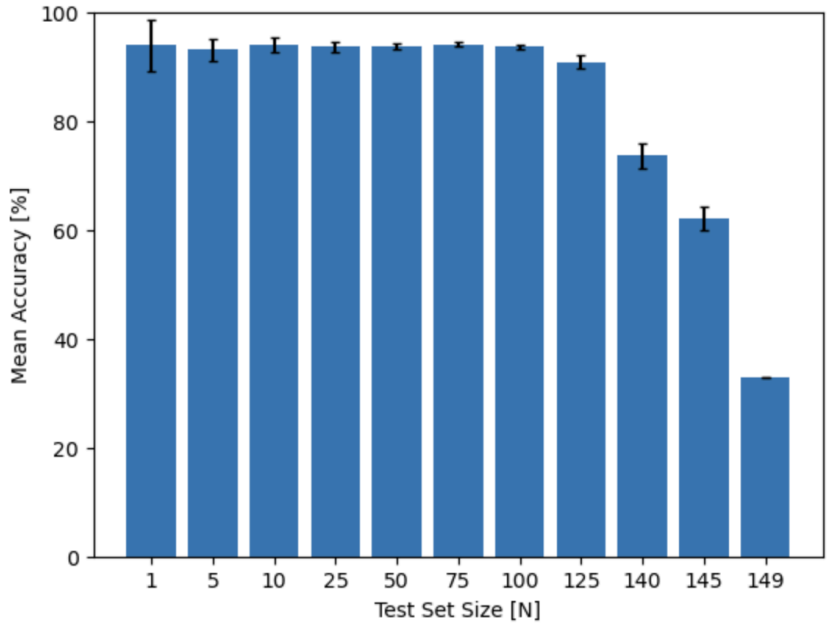

Figure 4a. The accuracy of the ID3 decision tree on the iris dataset.

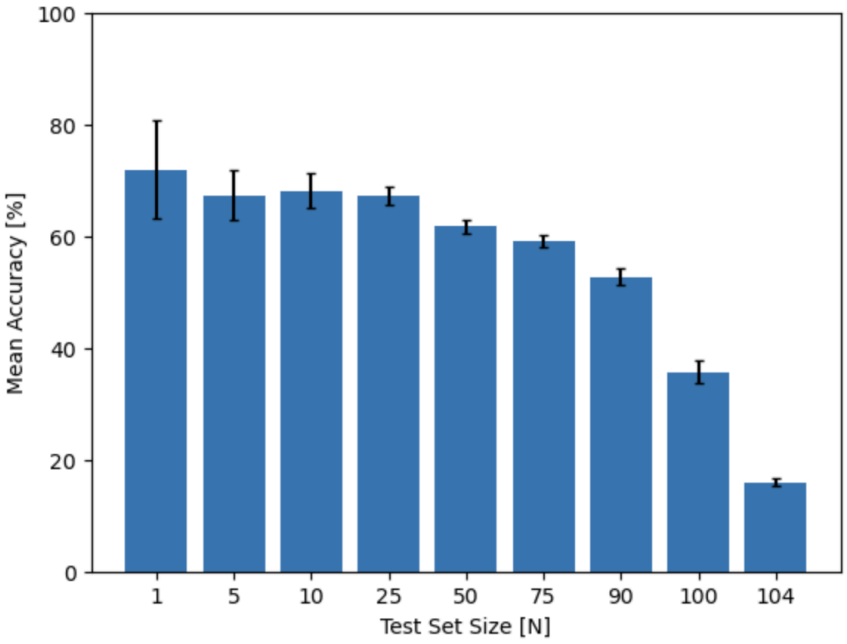

Figure 4b. The accuracy of the ID3 decision tree on the iris dataset.

Analysis & Conclusion

Based on the charts displayed above, it can be inferred that the split of the training v. test sets determines how correct the classification. When the training set was on the larger side, the algorithm seemed to perform better and when the training set was on the smaller side, the algorithm did not perform as well. The iris data set processed better results in comparison to the cancer data set. Since the iris data set has more data points than the cancer data set, it can also be inferred that the algorithm works better with the more data points it is provided overall. The

cross-validation we have conducted shows how well the ID3 decision tree works at predicting the best classifications when given a varied amount of data points.

Sources

Brownlee, Jason. “How to Implement the Decision Tree Algorithm from Scratch in Python.” MachineLearningMastery.Com, 11 Dec. 2019, machinelearningmastery.com/implement-decision-tree-algorithm-scratch-python/.
“Decision Tree in Machine Learning.” GeeksforGeeks, GeeksforGeeks, 15 Mar. 2024, www.geeksforgeeks.org/decision-tree-introduction-example/.
Rathore, Pratima. “Complete Guide to Decision Tree.” Medium, Medium, 27 Aug. 2020, iprathore71.medium.com/complete-guide-to-decision-tree-cee0238128d.
Saini, Anshul. “What Is Decision Tree? [A Step-by-Step Guide].” Analytics Vidhya, 18 Apr. 2024, www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/.
Sakkaf, Yaser. “Decision Trees for Classification: Id3 Algorithm Explained.” Medium, Towards Data Science, 12 Sept. 2020, towardsdatascience.com/decision-trees-for-classification-id3-algorithm-explained-89df76e72df1.