Jacob Anderson

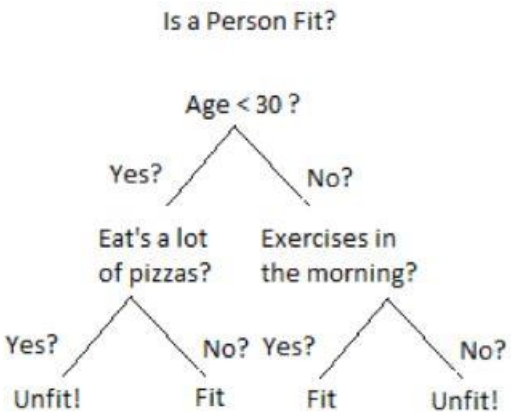
OLA 3 Report

CSCI 4350

Dr. Phillips

OLA 3

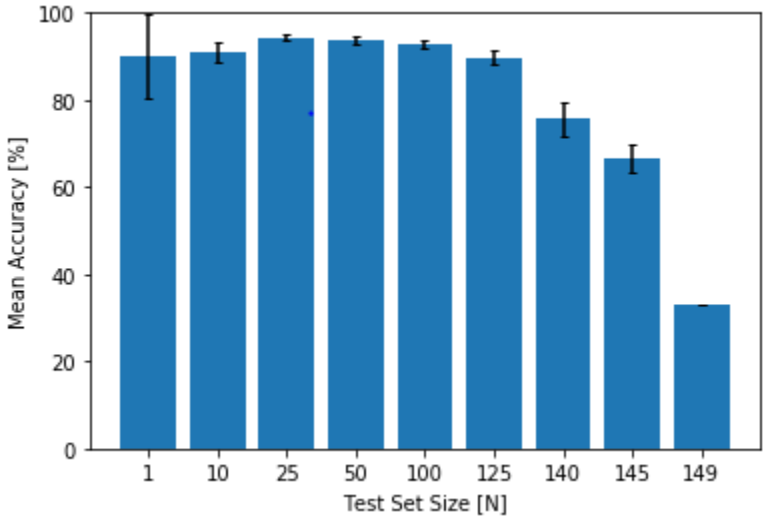
A decision tree is defined as a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements. If you look in figure one you will see an example of a decision tree. It examines variables such as age, eating habits, and exercise to determine whether someone is fit or not. In OLA 3 we were assigned to develop software that creates and learn an ID3 decision tree from labeled classification data. After doing this the software will then read in another set of data and be able to classify the information given in the testing data.



ID3 stands for Iterative Dichotomiser 3 algorithm. The algorithm was invented by a computer science data mining researcher by the name of Ross Quinlan. The algorithm works by taking in the probabilities of picking each classification in the set. After it gets the initial information it will then calculate the amount of uncertainty of each attribute, this is also referred to as the entropy. Once the entropy of each attribute is calculated, you then subtract that entropy of the attribute from the initial information, this is known as your information gain. You will pick the attribute that provides you with the most information gained. Once it selects and attribute it will then read in the data from each column of that particular attribute. Then it will calculate the amount of entropy between each value change in the column. The value change that has the best entropy will then be known as the split point. Using the split point you will create 2 child nodes and you will assign all the values that are less than the split point and the values that are greater than the split point to a different node. You will repeat this function recursively until the Information is 0. Once the information is 0 you will create a terminal node that determines the classification of the data.

For this project I created a Node class. This class defines each nodes, left and right child, split point value, the index of the split point, whether the node is terminal or not, and the classification of the node. The node class is used in a function called Build_tree(). In this function I define a new node and calculate its split point value, its information, and its category

index. Once It completes this it will take all the values from the data given and will split the values up by the split point and put then in either left or right child data. Then it recursively calls Build tree using left or right child data until the information equals 0. Once its information gets to 0, it sets the node to be a terminal node and assigns its classification. After the program builds the tree you take in other information to cross validate it to check for correct answers. To do this I read in the set of testing data to a np array. After doing this I used a variable called current and set it equal to the root node. From there I examined if the current node was Terminal or not. If the node was terminal then the program will check to see if the classification of the testing data matches the classification of the terminal node. If it does it will add one to the amount of correct. If it doesn't match then it will move on to the next item In the test data array. If the node is not terminal it will check to see if the if the value of the data attribute is greater than or less to the node split point. It will then assign current to either the left or right child node.



If you look at the graph above, you will see the results of the id3.py program. The data shows the mean accuracy of total number of classifications answered correctly based on N which represents the total number of test data that is examined. As you can see as N increases the accuracy begins to decrease. This is likely because the more data that it examines the more likely it is to answer incorrectly and the more data that it examines the harder it is to find the variables that connect to each other to find the answer.

While programing this project I kept running in to a couple of problems. For one, I was unable to run my code using "./id3.py" and had to run it using "python3". Another problem that I ran into was that on some instances of testing the test data I kept getting a Unlocal bound error and that my variable was not assigned before reference.

Works cited

Maurya, Siddharth. “Decision Trees - Introduction (ID3).” *Medium*, Towards Data Science, 4 Nov. 2019, https://towardsdatascience.com/decision-trees-introduction-id3-8447fd5213e9.