John Geddes

CS 460G Machine Learning

Homework 1

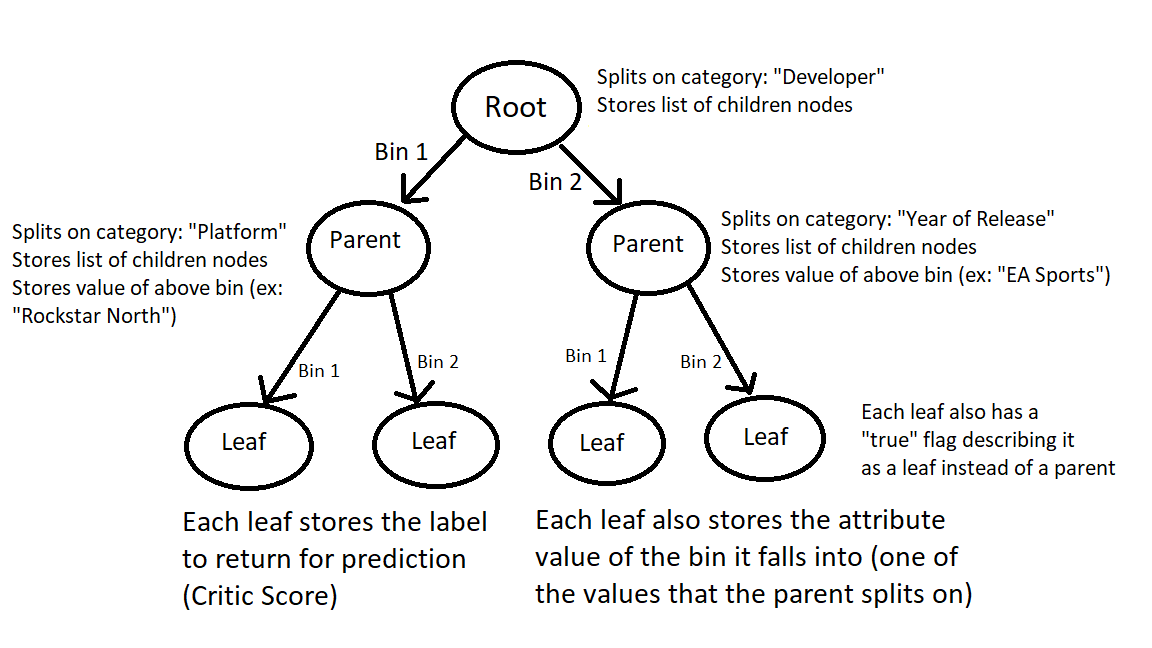
February 7, 2018

Implementation Decisions

I decided to use the library *numpy* to use the *log2* function and the *unique* function. The *unique* fuction takes in a list and returns a list of every unique value. This was especially useful when finding the most common labels in a bin and calculating entropy of a node.

I used a 3D list as the data structure of my program. The top layer was a list of all bins. The second layer was a bin’s list of data rows. And the third layer was a row’s list of attribute values. This didn’t pose problems with the synthetic data because I could split the continuous data and map them to the same indices as the attributes in the list. But when the program built the tree for the video game sales, the attributes were strings instead of continuous numbers. I solved this problem by making a list of all attribute values (ex: list of all Developers) and finding the index of the current bin using that list.

The 3D list is just how I store the data as I build the tree. The final tree is an actual tree of classes called *Node*. Each Node can either be a parent with children or it can be a leaf with no children. If the Node is a parent with branches, it stores an attribute that corresponds to the column in the data that the branch splits on. If a parent Node has a parent itself, then it also stores a value that corresponds to the branch between this Node and its parent above. If the Node is a Leaf, it has a flag changed to *true* to indicate it’s a leaf. A leaf doesn’t have a list of children nodes because it doesn’t have any children. A leaf also stores the value of the bin that it falls into.



Amazing diagram of a tree (totally not made in MS Paint)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Data Set | synthetic-1 | synthetic-2 | synthetic-3 | synthetic-4 | Video-Game-Sales |
| Training Error | 0.0 | 0.115 | 0.165 | 0.385 | 0.742776343293 |

The plots don’t have the sample background colors because I spent so much time debugging the ID3 algorithm. I hardly had time by Wednesday night to finish the training error and begin plotting sample data.

