David Wolff

Machine Learning

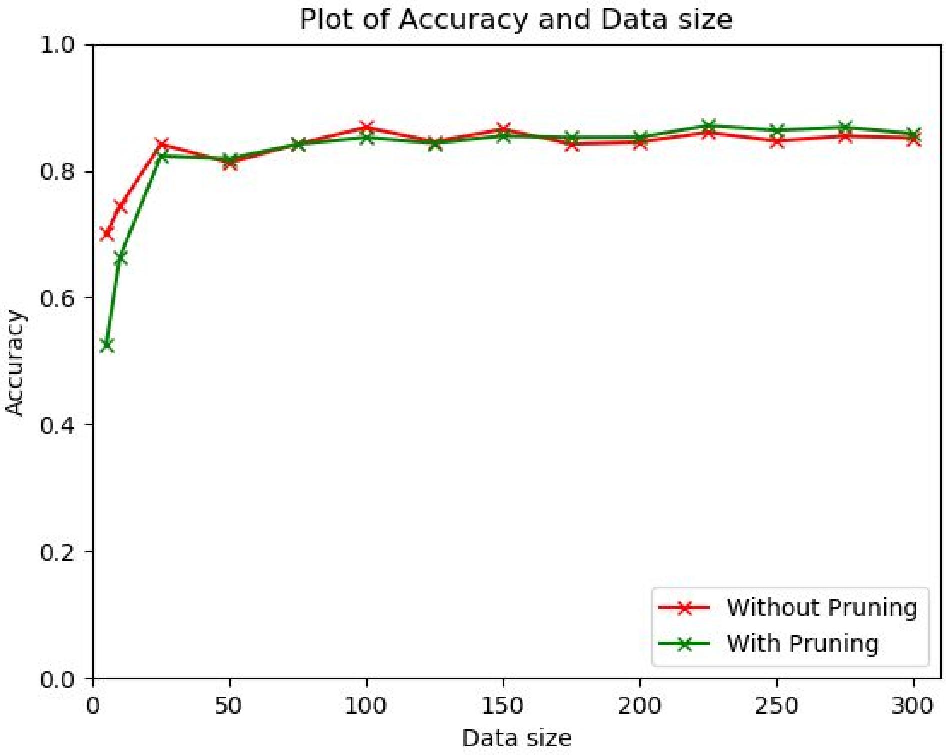
EEGS 349 - PS1

Downey, Spring 2018

Implementation of ID3 Decision Tree

1. Nicolas Finkelstein is my partner.
2. We did alter the Node data structure. We added three more classes to the structure than what was given. We added in self.name to store the attribute title, self.parent for reference to the node that has the current node as a child (this is for searching), and self.mode which gave us the most common attribute of its children (extremely useful when parsing). We used self.label to store the classification value if the node was a leaf (set to None otherwise).
3. For missing attributes, we made sure to run our preprocessing on the examples before we ran id3 and pruning. This replaces the missing attributes with the mode of the dictionary, given a dictionary with missing values. If there is a tie, then the first attribute that tied will be used to replace the missing value. We chose this method because it was the fastest to implement as well as only giving us 2 values to work with instead of 3.
4. The way we performed pruning was a depth first search through the tree. Given the root node of the tree (the first attribute), we fill a stack with its children and their children's' children until there are no children of the current node. From this, we acquire all nodes of the tree. From here, we can check if pruning that leaf will increase the accuracy of the tree given the validation data. If we prune that node, we append the parent node into the stack so that our function can check if we want to prune that node (if it has two leaf children).

5. The figure of the comparison of with pruning and without pruning can be seen on the right.



a. The general trend of both of the lines show that as the size of training set increases, the accuracy increases. At about ~30 examples, the increase in accuracy begins to taper off. This makes sense because the greater the training set, the more 'information' goes into creating the tree and therefore the more 'patterns'/relationships between the attributes we can find. Since the pruned tree modifies the original tree, it's accuracy also increases as the training set size of the originally tree goes up.

b. Yes, the data set size has an affect on how advantageous our pruned tree will be. Our plot above shows how the original decision tree's accuracy is higher than that of the pruned tree up until ~50 examples are used. At bigger data set sizes, the original tree becomes susceptible to overfitting and thus losing accuracy whereas the pruned tree takes care of the overfitting. Pruned trees are also smaller than unmodified ID3 decision trees because they contain less nodes and therefore perform quicker.