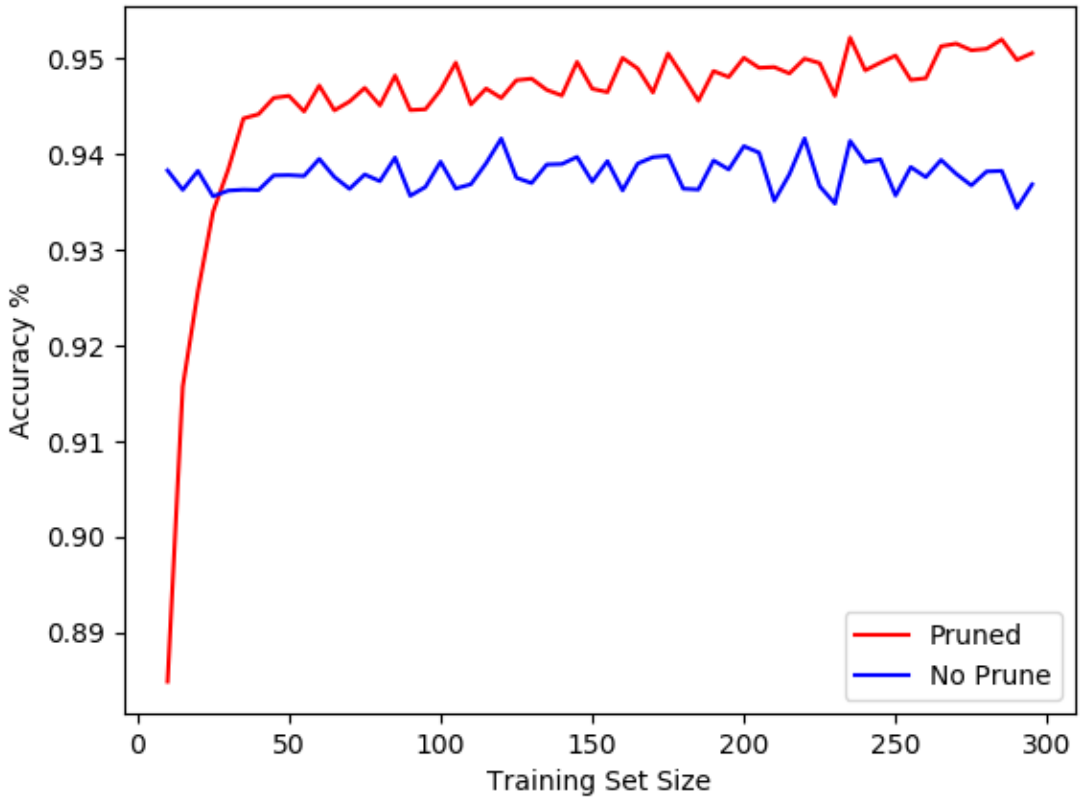
Eric Yang
10/16/18
MSAI 349

Decision Tree/Pruning Problem Set #1

1. I did this solo, without any other student
2. I added a few object variables to track a couple of things:
    a. Examples- list of examples that are at this node and under. Very useful for the recursive function to reference this as the worklist. And need to refer to this for pruning.
    b. isLeaf- Track if the node is a leaf or not as a stop condition for recursive calls
3. For missing attributes, I took in '?' as a potential value for the attributes and grouped examples that had missing attributes together. I didn't want to remove examples that were missing attributes because the dataset was limited to begin with.
4. My pruning algorithm started at the leaf nodes and went up the tree (recursively) and identified each of the nodes that pointed to only leaf nodes as potential nodes to be pruned. To evaluate whether it should be pruned or not, I would first calculate the accuracy of the tree as is with the validation set, then carry through with the prune setting the MODE of the examples at that node as the value, and calculate the accuracy of the tree after the prune with the validation set. If the calculated accuracy of the pruned version of the tree was LESS than the accuracy before the prune, then I would un-do the prune. This evaluation goes from the bottom of the decision tree up, so it re-evaluates nodes that are further up after its children have potentially already been pruned.
5. Accuracy Plot of pruned and unpruned trees:



a. The two trendlines generally get higher as the training set is larger, which makes sense because the more data it has to train on, the better the tree will be and more accurate.
b. The pruned trendline hits a plateau at a training size around 40. And it has a steep ramp up before that, but it underperforms the unpruned trees with training sets below 40. This makes sense because with a small training set size, it is easy to overprune, since using the mode of the outcome of few samples is not accurate.