Ikhlas Attarwala / 10-17-18
Machine Learning / MSAI 349
Doug Downey / Problem Set 1

1) Vamsi Banda and I were in a group. (Worked together. Code uploaded by us is the same)

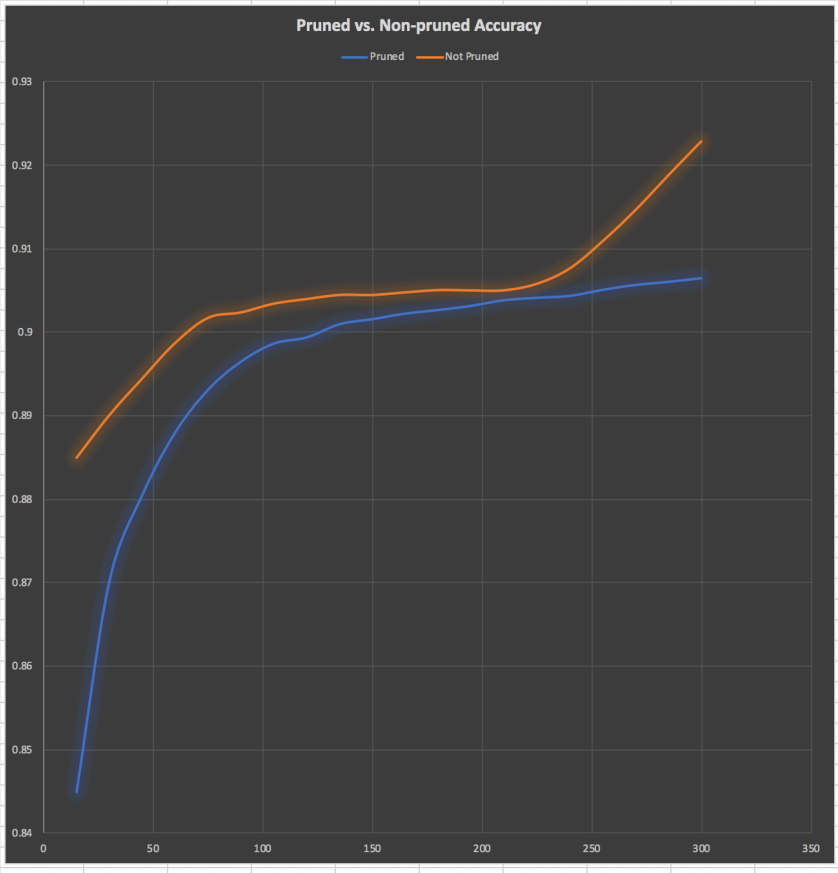
2) We altered the Node by adding *set_lable()*, *get_lable()*, and *set_children()* definitions (note that *set_lable()* is also referred to as a mutator method). Member variables are made private, and *set_lable()* controls changes to that variable (similar to *set_children()* on child) while *get_lable()* returns that private member variable's value.

3) We considered all missing attributes as a separate category. Vamsi said that missing attributes can have their own "value", which is to say that they can still have influence or worth in the fact that they are "missing".

4) Simply put, to reduce overfitting, we chose to prune the tree correctly to preserve greater accuracy in results. When post-pruning, we checked whether deletion of a leaf increased accuracy on validation, and if it did, we would keep the tree "structure" without the recently deleted leaf and move on to the next node. If deletion of that leaf decreased accuracy, we would instead go back and keep that node/leaf in our tree. This method is more accurate than pre-pruning and allows us to remove non-significant children (not real children) from the tree.

5) Table showing Pruned vs Non-pruned accuracy on **house_votes_84.data**

	X-AXIS	Y-AXIS (ACCURACY AS A DECIMAL [0 - 1])	
#s	TRAINING SET SIZE	PRUNED ACCURACY	NOT PRUNED ACCURACY (OVERFIT)
1	15	0.844862385	0.885045872
2	30	0.870137615	0.890137615
3	45	0.880489297	0.894525994
4	60	0.887981651	0.898761468
5	75	0.893100917	0.901761468
6	90	0.89648318	0.902415902
7	105	0.898636959	0.903460026
8	120	0.899392202	0.904013761
9	135	0.9009684	0.904505607
10	150	0.901577982	0.904504587
11	165	0.902243536	0.904812344
12	180	0.902675841	0.905084098
13	195	0.903175723	0.905045872
14	210	0.903846658	0.905058978
15	225	0.904140673	0.905834862
16	240	0.904369266	0.907672018
17	255	0.905089045	0.910922828
18	270	0.90567788	0.914653415
19	285	0.906050217	0.918802511
20	300	0.906486239	0.922862385



a. As the training set size increases up to 300 (largest set size ran), both accuracies for pruned and non-pruned trees increase. This makes sense because the greater number of trials you test upon, the more accurate your expected results should be.
b. The advantage of pruning is clearly depicted above by how the blue line (pruned tree) ends on a more linear and stable growth over trials, almost as though it were reaching a limit or threshold, and unlike the orange line (non-pruned tree), avoids overfitting the results. It makes more sense to model after because we have reduced noise, can fit new and additional data with less consequence, and even predict future observations more reliably.

Fin.