Chaitra Subramaniam: cso330

1. 0.5 points) Which other student, if any, is in your group? (either names or netIDs is fine)

   Gabe Rojas-Westall: gwr744

2. (0.5 points) Did you alter the Node data structure? If so, how and why? (2 sentences)

Yes, we included an attribute to track which attribute it is on easily. We added children to track the different children nodes as well as a majority to track the mode class of the examples that was split at that node (0/1 or Democrat/republican etc)

3. (1 point) How did you handle missing attributes, and why did you choose this strategy? (2 sentences)

We treated missing attributes as it's own value such as y/n. We thought that missing values do carry certain information that might be lost if we don't account for them in more detail.

4. (1 point) How did you perform pruning, and why did you choose this strategy? (4 sentences)

We did reduced error pruning using the majority attribute of the node class. We used depth first search to traverse the tree, starting from the nodes at the bottom of the tree, and we compared accuracy for each node without its children to the original accuracy of that node with its children. If the pruned node returned a better accuracy, then we dropped the children for that node from the entire tree and set the value of that node to be the majority attribute that we added for each node because that was the class that was most prevalent among its children. We chose this strategy as it is simplified and fast to use and we already have a function to find accuracy (Test()) to use for pruning.

5. (2 points) Now you will try your learner on the house_votes_84.data, and plot learning curves. Specifically, you should experiment under two settings: with pruning, and without pruning. Use training set sizes ranging between 10 and 300 examples. For each training size you choose, perform 100 random runs, for each run testing on all examples not used for training (see testPruningOnHouseData from unit_tests.py for one example of this). Plot the average accuracy of the 100 runs as one point on a learning curve (x-axis = number of training examples, y-axis = accuracy on test data). Connect the points to show one line representing accuracy *with* pruning, the other *without*. Include your plot in your pdf, and answer two questions:
   a. In about a sentence, what is the general trend of both lines as training set size increases, and why does this make sense?
   b. In about two sentences, how does the advantage of pruning change as the data set size increases? Does this make sense, and why or why not?

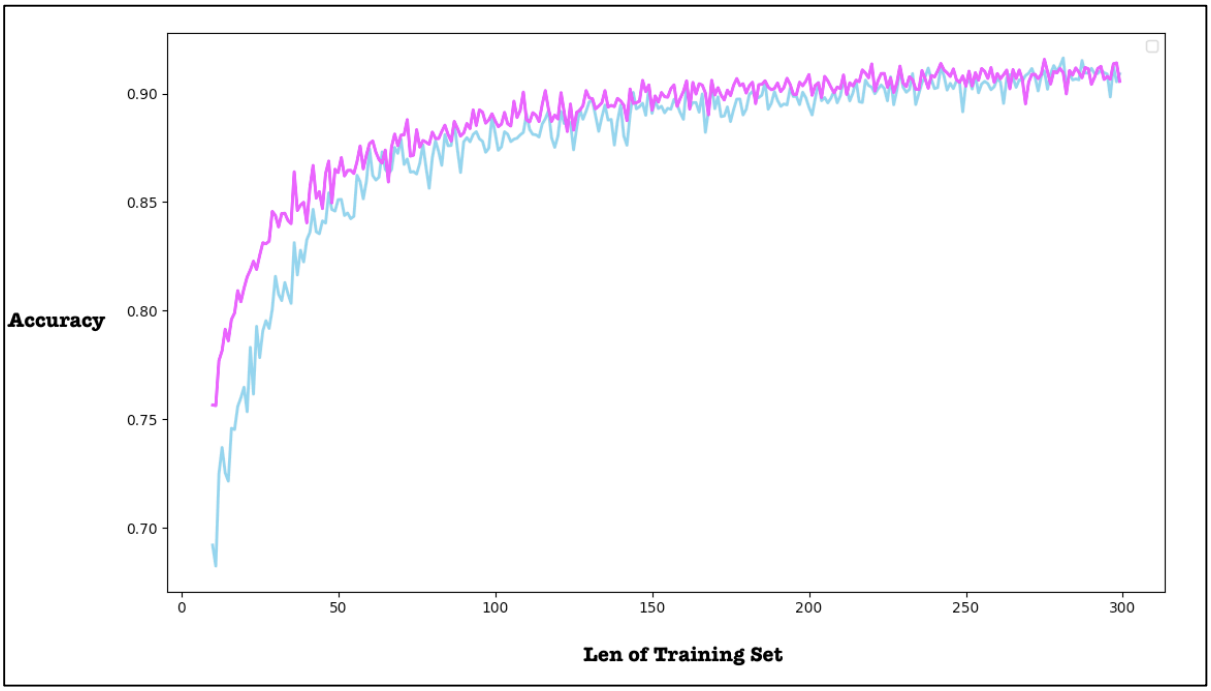

Figure 1: Without Pruning (Magenta) and With Pruning (Blue)

a. As we pass more examples into the training set, the accuracy for classifying any one given example increases because the tree is trained using more examples.
b. Pruning becomes more advantageous as the tree gets larger. This makes sense because with a larger dataset, there is more noise so pruning can help remove this noise significantly and have more impact in larger datasets.