CS 373 Fall 2019
Instructor: Romila Pradhan
Byoungjin Oh
0030411229
I collaborated with (Kate Lee, Young Suk Moon). I affirm that I wrote the solutions in my own words and that I understand the solutions I am submitting.

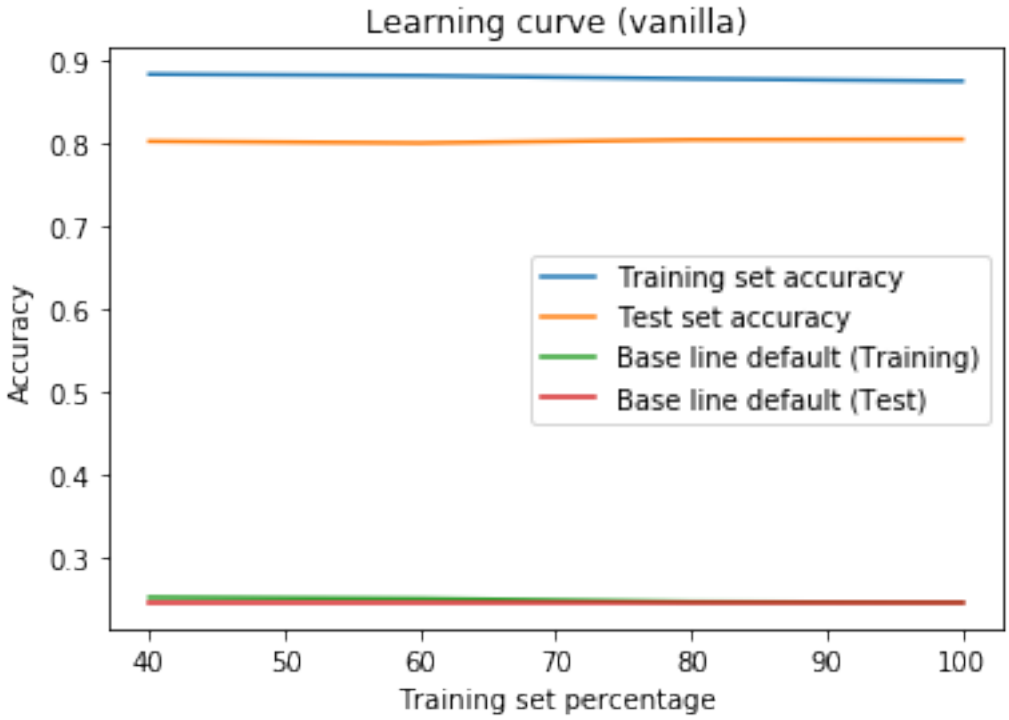
CS373 Homework 3

Due date: Wednesday, October 30, 11:59pm

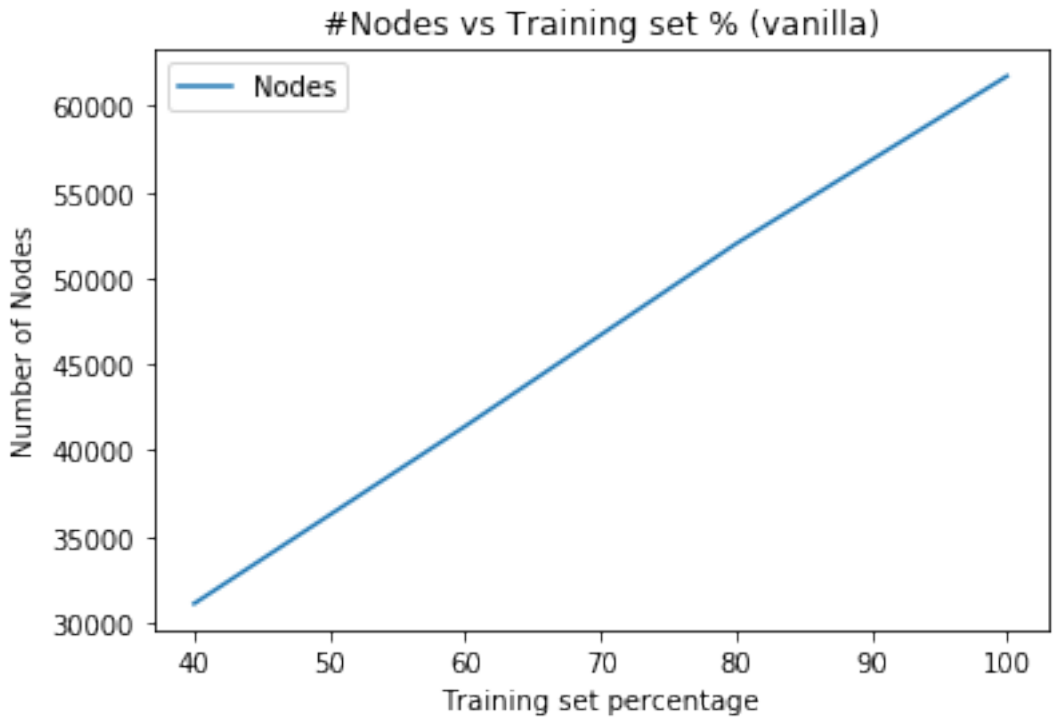
Part 2 Analysis

1. (**7 points**) For the full decision tree (vanilla), measure the impact of training set size on the accuracy and size of the tree.
   Consider training set percentages {40%,60%,80%,100%}.

   - Plot a graph of test set accuracy and training set accuracy against training set percentage on the same plot. Also include a line for the baseline default error that would be achieved if you just predicted the most frequent class label in the overall data (100% of the train data).



- Plot another graph of number of nodes vs training set percentage.

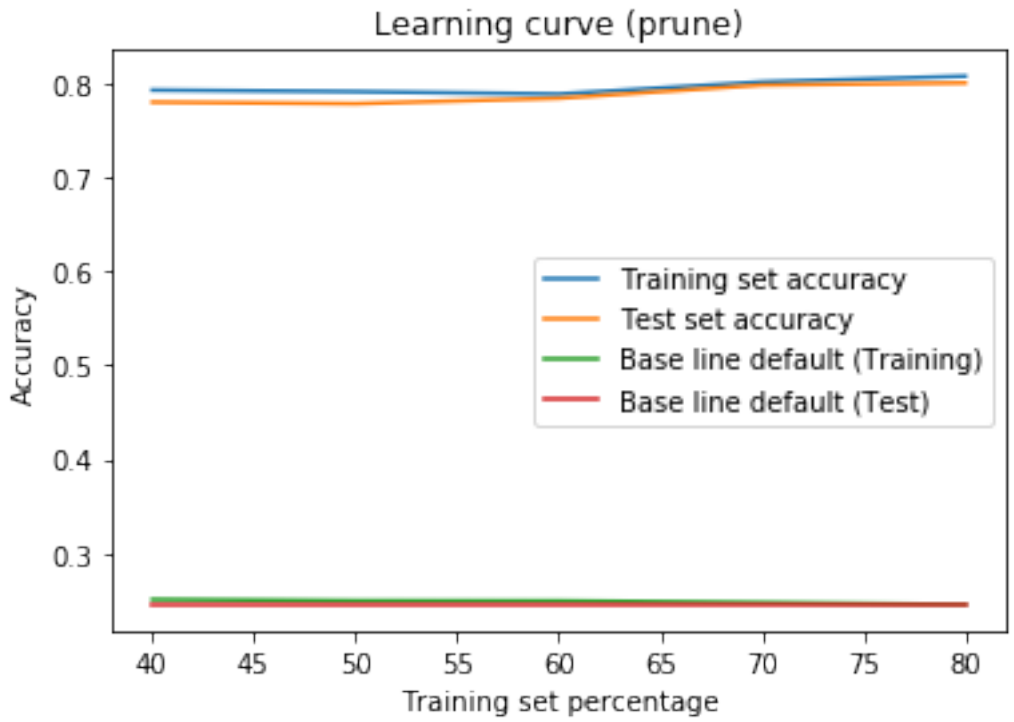


- Discuss the results (e.g. how is accuracy impacted by training set size and how it compares to just predicting the dominant class) in a few sentences. For each training set size, is the decision tree overfitting?

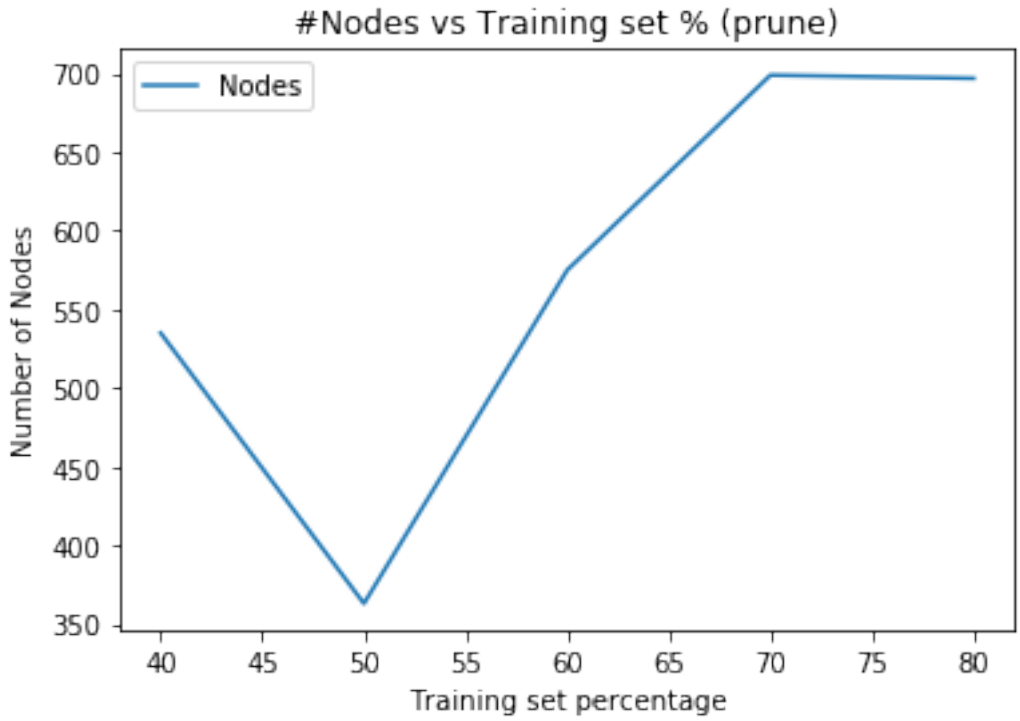
  As the training set size increases, the number of nodes increases, the training set accuracy decreases and the test set accuracy is not very changed. For each training set size, the decision tree is overfitting because the training set accuracy is way higher than the test set accuracy.

2. (**8 points**) Repeat the above analysis for the pruning case (prune).

- Again, consider values of training set percentage: {40%,50%,60%,70%,80%}. The validation set percentage will remain 20% for all the cases. You will use the validation set when deciding to prune.
- Plot a graph of test set accuracy and training set accuracy against training set percentage on the same plot. Also include a line for the baseline default error that would be achieved if you just predicted the most frequent class label in the overall data (100% of the train data).



- Plot another graph of number of nodes vs training set percentage.



- Discuss the results (e.g. how is accuracy impacted by training set size and how it compares to just predicting the dominant class) in a few sentences. For each training set size, is the decision tree overfitting?

As the training set size increases, both training set accuracy and test set accuracy increases slightly and the number of nodes is not linear. For each training set size, I would say the decision tree is not overfitting. Because, even though the training set accuracy is still higher than the test set accuracy, they are almost same and the difference between them is much less than the difference of vanilla accuracies.

Part 3 Theoretical questions

1. (**2 points**) Will ID3 always include all the attributes in the tree?

   No, ID3 will not always include all the attirubtes in the tree. While ID3 is iterated recursively, the attribute which is used for the node will be removed to the sub-set that will be used to get the sub-node.

2. (**2 points**) Why do we not prune directly on the test set and use a separate validation set instead?

   Test set should be used only for testing the final model in order to confirm the actual accuracy. If we prune on the test set, we do not know what the actual accuracy is.

3. (**3 points**) How would you handle missing values in some attributes? Answer for both during training and testing.

   The missing values can be handled assuming them as the most common value for both cases. So for during training, the missing value will be assumed most common value regarding to the attribute and for during testring, go to the most common branch label.

4. (**4 points**) How would you convert your decision tree from a classification model to a ranking model (i.e., how would you output a ranking over the possible class labels instead of a single class label)?

   I would change the leaf node to store the frequnce of the values of class label instead just represents with most common value of the class label and then sort the values according to its frequncy with descending order and return the ordered values which starts with the highest ranking descends to lower rankings.

5. (**4 points**) Consider the case where instead of binary values, the class label has continuous values. How would you adapt your decision tree to predict the class value of a test instance? (4 points)

   I would use the threshold. Firstly sort the instances according to the value A, find the best split point calculating the entropy of the candidate points which each has different subtrees and then the best split point will be threshold. Build a node A with branches according to the threshold.

   Ex) Attribute A, threshold B, the branches are the value of A is less than B and the value of A is bigger than or equal to B.

   Lastly, during testing, get the value of the attribute A, and check if the value is satisfied to the branches.

Part 4 Extra Credit: Max Depth

2. Tune your maxDepth value on the validation set. Briefly discuss how you did this and what maxDepth value you selected. Plot a graph of at least 10 maxDepth values you tried on the validation set (including the one you selected), and compare with the performance on the validation set.

   I selected maxDepth value which has locally maximum validation set accuracy at the first area (finding the point the validation set accuracy decreases at first time). The maxDepth I selected is 18.