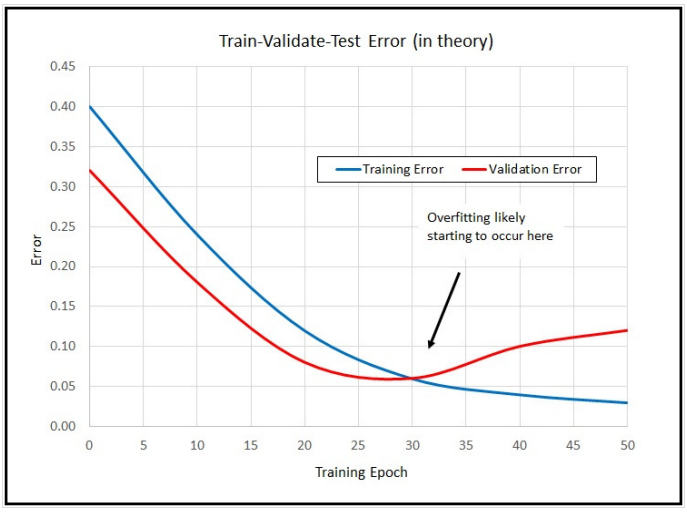
Akash Lankala
PUID: 0027710383
CS 373
Homework 3
October 30th, 2019

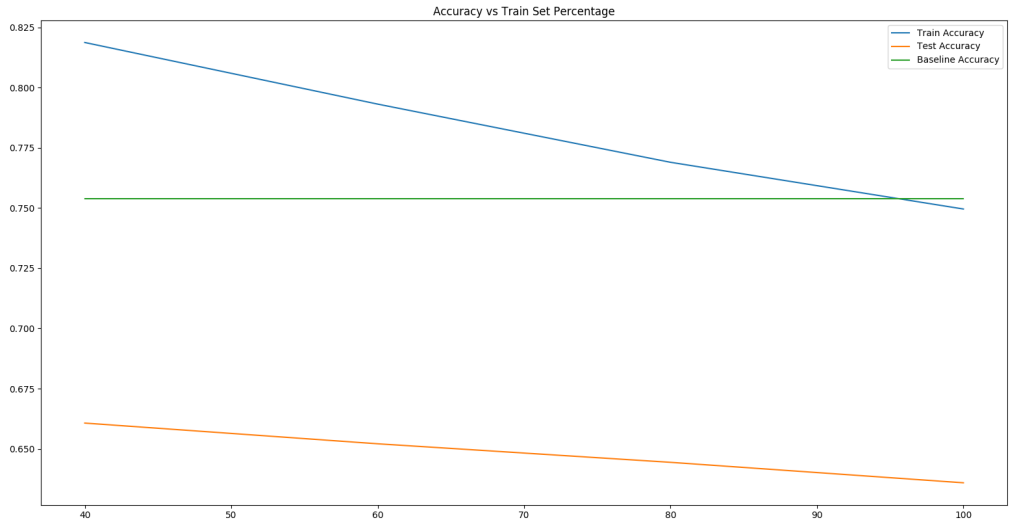
Late Days Used: 0
Collaborated with: no one

Part 2 Analysis

1. 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%}.

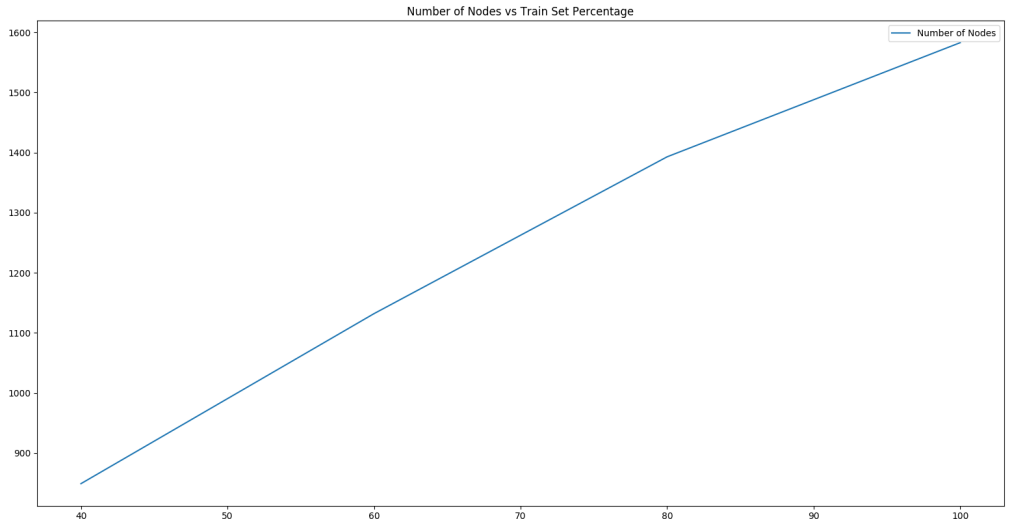


- Graph of test set accuracy and training set accuracy against training set percentage on the same plot.



Akash Lankala
PUID: 0027710383
CS 373
Homework 3
October 30th, 2019

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

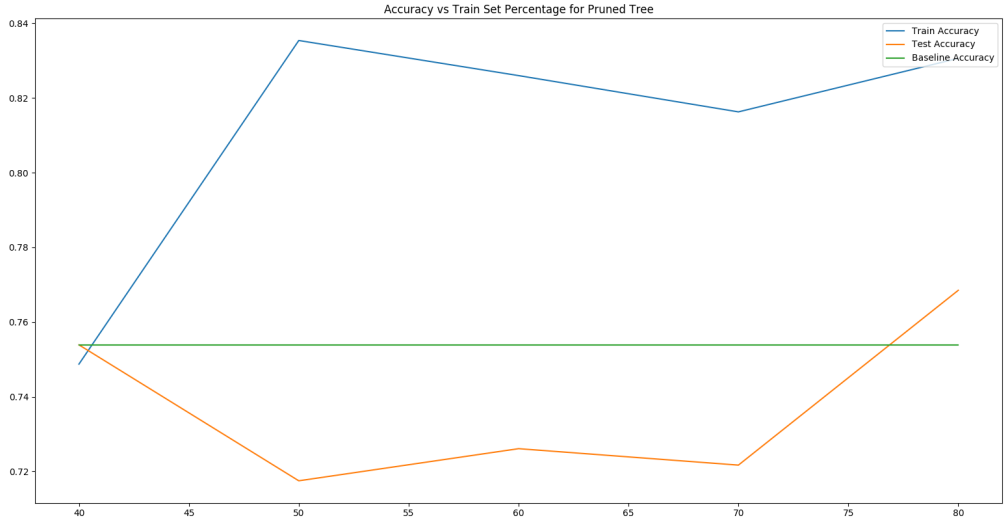


- Discuss the results (e.g. how is accuracy impacted by training set size and how it compares to just predicting the dominant class) in a few sentences. For each training set size, is the decision tree overfitting?
    - The accuracy on the test set decreases since the model overfits as the training dataset size increases since we are not validating in this case.
    - Since we are dealing with binary classification and there is a case of imbalanced classes in the dataset, the baseline accuracy is high (~75%) as compared to the ID3 accuracy.
    - Yes, the model fits for each of the training size. This can be seen in the plot as the test accuracy decreases with an increase in the train dataset size.
    - The number of nodes in the graph increases with an increase in the training dataset size.

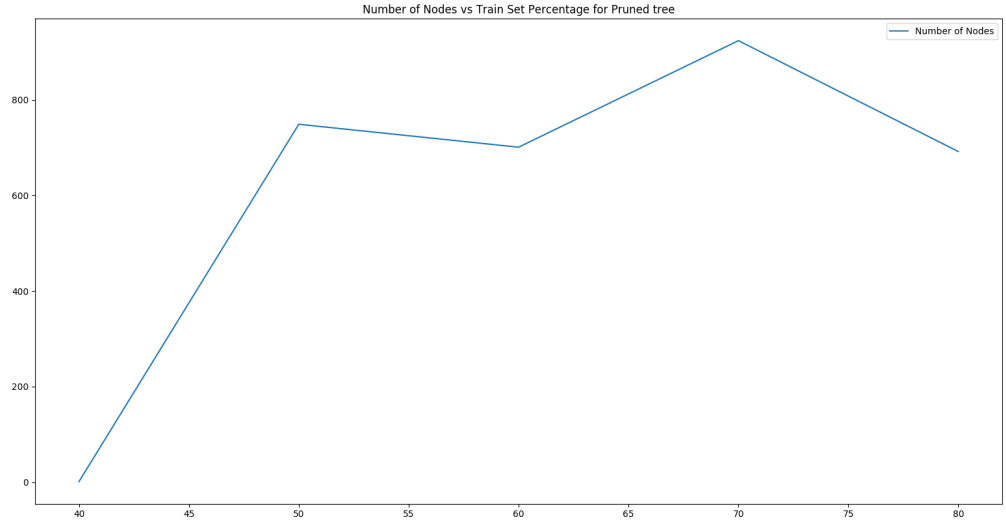
Akash Lankala
PUID: 0027710383
CS 373
Homework 3
October 30th, 2019

2. Repeat the above analysis for the pruning case (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.



Akash Lankala
PUID: 0027710383
CS 373
Homework 3
October 30th, 2019

- 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?
    - The accuracy on the test set fluctuates around the baseline accuracy since we are validating alongside training and are trying to prevent overfitting.
    - Since we are now cross-validating our model, when the training dataset size goes above a certain threshold (such as >70%), we receive better accuracy than baseline.
    - Through pruning, we are preventing overfitting by decreasing the size of our tree.

Akash Lankala
PUID: 0027710383
CS 373
Homework 3
October 30th, 2019

Part 3 Theoretical Questions

1. Will ID3 always include all the attributes in the tree?

No, ID3 will not necessarily include all the attributes in the tree. In this example, we can classify with just attribute A. So, the decision tree will not have attribute B since it will check if A=1, Class=k and A=0, Class=L.

A	**B**	**K**
1	1	K
1	0	K
1	1	K
0	1	L
0	0	L

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

We cannot drop the cross-validation set and use the test set for pruning because this would overestimate how good our model is. Using the test set, our model will optimize it according to it and show better results that are not real.

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

We can deal with missing values by:

- ⇒ Using the most common value of that attribute in data.
- ⇒ Using the most common value of that attribute in data with the same class label (at training time).

Akash Lankala
PUID: 0027710383
CS 373
Homework 3
October 30th, 2019

$\Rightarrow$ Assign fractional count instead of majority vote on value.

For the testing process, we can store the attributes' values using one of the above techniques and fill in the missing values for the test set.

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

For converting the decision tree from a classification model to a ranking model, we can traverse the complete decision tree instead of the single path as done in ID3. In preorder traversal, we can assign the edges as n or 0 (where n is a function of current depth, which deceases with an increase in depth) based on if our sample contains that property or attribute up to the leaf nodes.
We can then perform another tree traversal and add the weights from root to leaf along each path and assign the sum to the leaf nodes. Based on this, we can sum up the weights for leaf nodes which belong to the same class and normalize them to get the required probabilities of each class.

5. 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?

We can use thresholds or ranges to get Boolean tests or the range variables. We can split the continuous ranges based on the idea that the split changes the proportion of labels in the children and each child should have a higher proportion of one of the labels. In essence, we go over split points with different labels, and then find the one with the greatest information gain.
The split based on horsepower is an example of continuous range decision tree.