MSAI 349 HW1

Wentao Yao, Preetham Pareddy, Hualiang Qin, Bodhisatta Maiti

❖ **(1.0 points) Did you alter the Node data structure? If so, how and why?**

Yes, class attributes that have been added to the Node data structure are *: self.is_leaf, self.attribute,* and *self.best_edge*, and functions: *set_leaf(self, label)* and *add_child(self, label, node, best_edge)*.

- *is_leaf:* is used for detecting if the leaf of the tree is reached.
- *attribute:* is used for recording the attribute for the current node.
- *best_edge:* is used for the missing attributes which will be explained in detail in the next question.
- *set_leaf(self, label):* is used for setting *is_leaf* to True, and set the label
- *add_child(self, label, node, best_edge):* is using for adding children to the current node.

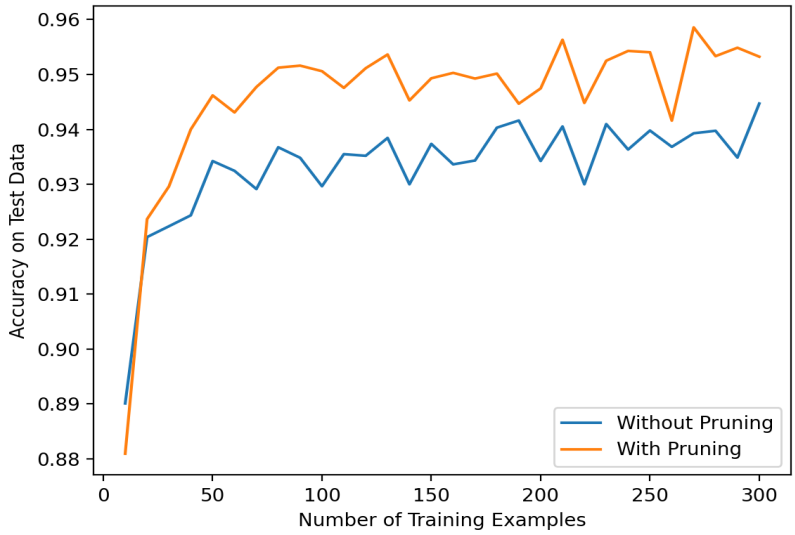
❖ (1.0 points) How did you handle missing attributes, and why did you choose this strategy?

- When building the tree, the missing attributes are treated as a special attribute value "?". Also, a class attribute, *best_edge,* is set up for the node. *best_edge* will record the most common value that attribute has. When making the prediction, if there is a missing attribute, and there is no "?" as the available attribute value(or I mean edge), then *best_edge* will be used as the alternative.

❖ (1.0 points) How did you perform pruning, and why did you choose this strategy?

- Reduced error pruning because it is a pruning method that can be easily understood and implemented.

❖ (2.0 points) Include your plot in your pdf, and answer two questions:



1. What is the general trend of both lines as training set size increases, and why does this make sense?

   The general trend of both lines increases as the training set size increases in the first hundred. After that the line becomes flat or we can say the accuracy does not change much. The reason behind it is that with a small training set, any outlier can affect the structure of the decision tree a lot, as the training set keeps increasing, the accuracy becomes more stable. Also, the tree learns more initially and hence the accuracy has a steep line, but after a threshold it flattens because the rate of learning decreases as the training size increases

2. How does the advantage of pruning change as the data set size increases? Does this make sense, and why or why not?

   No, there is no obvious increment of the advantage of pruning. This makes sense because the pruning method is used for pruning the overfitting edges and after a threshold of data size, the overfitting does not change much and therefore the advantage because of pruning is not changing much either.

❖ (Optional 1.0 points) Use your ID3 code to construct a Random Forest classifier using the candy. data dataset. You can construct any number of random trees using methods of your choosing. Justify your design choices and compare results to a single decision tree constructed using the ID3 algorithm.

The Random Forest classifier has been implemented in the random_forest.py file. It provides stable 73% accuracy on *candy.data*, compared to a single decision tree which only provides 67% accuracy.