EECS 349 HW2

Boyu Xu
bxp6650

1. Teammates: Boyu Xu(bxp6650), Zhanghexuan Ji(zjv5649), Daowei Li(dlx4580)

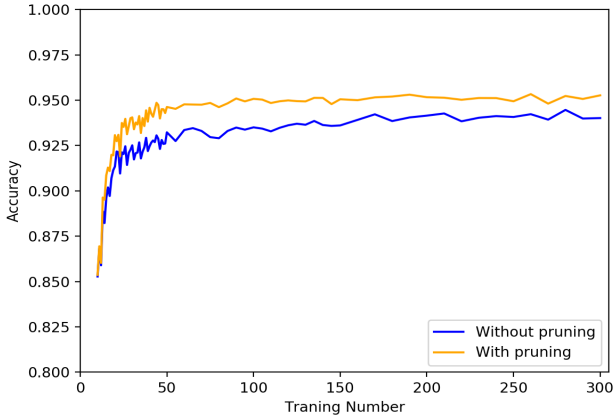
2. Yes, we add 3 attributes into the Node data structure.
   The first is “isleaf”, a Boolean bit, using 0 and 1 to denote if the node is the leaf node. The default value is 0.
   The second is “parent”, to record node’s parent node, the default value is none.
   The third is “train_num”, the data number for training of the leaf node.

   Record “parent” and “isleaf” condition is helpful for the construction of tree and also the pruning procedure. The train_num can be easily used for evaluate() and prune().

   3. When it comes to “missing attribute” as “?”, we simply assign it the value that is most common among training examples at current node n. It’s the most traditional ID3 method, the value that is most common does show the preference. Also by handling missing attribute this we can make the difference more obvious and assure accuracy.

4. We perform the pruning procedure by using post-order traversal of the constructed tree. For each leaf node, we compare the accuracy of the pruning tree and the origin tree, if the pruning one has higher accuracy, we did the pruning procedure and continuing traversing the tree, for each node we check the status. After the whole post-order traversal procedure, the tree has been pruned completely. This method has the effect that any leaf node due to coincidence or bias in the training set is more likely to be pruned since these coincidences are unlikely to appear in the validation set. And this method is also easy to implement and with a quite good efficiency. It only requires one transverse of the tree and two accuracy calculation for each node.

5.



a.

For both lines, with the training set size increases, the accuracy also increases. when training set is large enough, the learning curve will reach the plateau (with little fluctuation). It is because when training set size is small, the training data is very likely to have large bias and coincidence, when the training set size is large, the statistic distribution will make the data set nearly unbiased so that the decision tree can be more accurate.

b.

As the dataset size increases, the pruning tree has better performance than unpruned decision tree. This does make sense, The advantage of pruning will firstly increase as the data set size increase, and then remain at a constant level. This is because pruning can only handle overfitting and reduce some nodes that contain redundant attributes or the nodes that are generated due to random noise, when the dataset size is small, the decision tree itself is not accurate, then the pruning procedure cannot improve its performance, when the size of dataset is large enough, pruning procedure can handle the overfitting which is helpful for improve the performance.