

Problem 1

Part 1a Inducing the initial decision tree

$$Gain(SHAPE) = I\left(\frac{4}{7}, \frac{3}{7}\right) - \left[\frac{3}{7}I(1,0) + \frac{3}{7}I\left(\frac{1}{3}, \frac{2}{3}\right) + \frac{1}{7}I(0,1) \right] = 0.59$$

$$Gain(AGE) = I\left(\frac{4}{7}, \frac{3}{7}\right) - \left[\frac{4}{7}I\left(\frac{2}{4}, \frac{2}{4}\right) + \frac{3}{7}I\left(\frac{2}{3}, \frac{1}{3}\right) \right] = 0.02$$

$$Gain(WORTH) = I\left(\frac{4}{7}, \frac{3}{7}\right) - \left[\frac{3}{7}I\left(\frac{2}{3}, \frac{1}{3}\right) + \frac{4}{7}I\left(\frac{2}{4}, \frac{2}{4}\right) \right] = 0.02$$

To select the feature with maximal information gain, we select SHAPE as root node. After selecting SHAPE, the nodes with SHAPE = C, S and T (labeled as the majority class “-”) become leaf nodes. When SHAPE=E

$$Gain(AGE) = I\left(\frac{1}{3}, \frac{2}{3}\right) - \left[\frac{1}{3}I(1,0) + \frac{2}{3}I(0,1) \right] = 0.92$$

$$Gain(WORTH) = I\left(\frac{1}{3}, \frac{2}{3}\right) - \left[\frac{2}{3}I\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{1}{3}I(0,1) \right] = 0.25$$

So, this time we will select AGE. The decision tree is as Fig 1 shows.

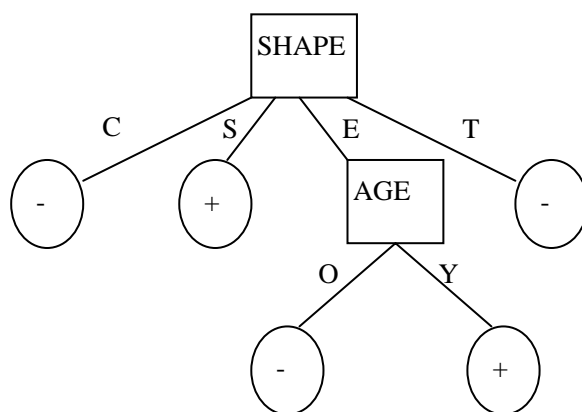


Fig. 1

Part 1b Pruned Tree

The accuracy on the original tree is **0.6**. Using the greedy pruning algorithm given in the homework, the initial bestTree is the original tree and initial bestAccuracy is 0.6.

Then we go into the while loop, and first we prune the SHAPE node. The whole tree becomes a leaf node with class negative. Its accuracy is **0.6**. Now, bestTree = a tree with a leaf node labeled negative, bestAccuracy = 0.6.

By pruning the AGE node, we produce the pruned tree in Fig. 2. Its accuracy on the tuning set is **0.8**. bestTree = the tree in Fig.2, bestAccuracy = 0.8.

Then we go to the second iteration of the while loop, with currentTree = the tree in Fig.2. We prune SHAPE node, the accuracy is $0.6 < 0.8$. So progressMade = “false” and we stop the iteration.

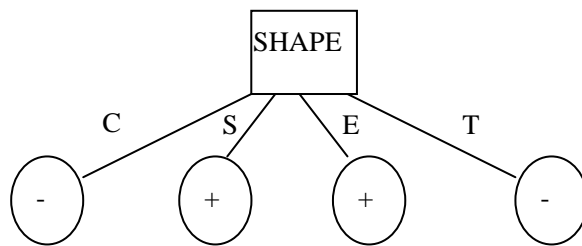


Fig. 2

So we choose the tree in Fig. 2 as our final tree.

Part 1c Estimating future accuracy

The testset accuracy on the original tree in Fig 1 is **0.6**, while the testset accuracy of Fig. 2's pruned tree is **0.8**. This example shows that pruned tree has a better performance than the original tree, which overfit the training set.