Anusha Prasad
CS-3600-A
Project 4A: Decision Trees

# Analysis

Question 6:

Recorded classification rate and tree size statistics for data sets:

- DummySet1:
  - Tree size = 3
  - Average classification rate = 1.0
  - Correct classification rate = 1.0

- DummySet2:
  - Tree size = 11
  - Average classification rate = 0.65
  - Correct classification rate = 0.55

- Car:
  - Tree size = 408
  - Average classification rate = 0.9440000
  - Correct classification rate = 0.95

- Connect4:
  - Tree size = 41521
  - Average classification rate = 0.757550
  - Correct classification rate = 0.75

Explanation:

For the first data set (i.e. DummySet1), the accuracy rate is 100% because the classification rate resulting from training was 1.0 (same as the correct classification rate). When examining the printed evaluated tree for the first data set, it appears small. This is because attribute 5 specified for the decision tree determines the label for each of the samples. Based on the smaller tree size (i.e. 3), there is an attribute (i.e 5) that will classify the tree correctly in a small number of training samples (i.e. 20).

For the second data set (i.e. DummySet2), the accuracy rate is slightly lower than the previous data set (proving that this data set was unable to classify elements of the tree as efficiently as the first data set). It appears that in the evaluated tree that is produced, there are

splits on 5 attributes across the 20 training samples. Since the amount of training data available was small, there were certain attributes used to classify the decision tree (in reality these attributes might not be proper deciding factors for the tree). This issue causes the tree to incorrectly split on certain attributes (i.e. not making the best decisions and losing accuracy).

For the third data set (i.e. Car), there appear to be a large number of permutations for the attributes of the data set (i.e. buying, doors, maint, persons, lug_boot, and safety; 360 possible permutations and 1728 possible combinations). This means that the decision tree is able to classify attributes of testing samples with great accuracy.

For the fourth data set (i.e. Connect4), the size of the data set is large (with 67557 instances and 42 attributes). The average classification rate for this data set is approximately 0.76 (a relatively low number), which indicates that the possible combination of attributes is not able to deterministically predict the winner of the game. For instance, if the board is extremely empty, there are a variety of possible future moves for each player (not optimal). The tree size for this data set is extremely large due to the possible permutations of various board positions for the player. Essentially, for this data set there are limited deciding attributes that classify the decision tree accurately.

Question 7:

Car data set:

An appropriately paired site that would display similar information would be Autotrader. This site allows users to filter through used cars, new cars, and certified pre-owned cars. Additionally, users are able to browse through make and model, with site locations. In the Cars data set, there is a decision tree produced based on attributes involving: doors, persons, safety, and buying price. Having this information coupled with finding the best prices available for certain cars (fitting the decision tree description in the Cars data set), a user is able to make the appropriate decision for their next car purchase (or sell their vehicle for a reasonable profit).

Connect4 data set:

The decision tree for the Connect4 dataset would be best for heuristic evaluation (based on a ‘win,’ ‘lose,’ and ‘draw’ state). These states are able to propagate upstream to give a resultant value with respect to cumulative values of available options for the decision tree. This process would be appropriate for providing an accurate representation of the effectiveness of move within the game (compared between the players). This data set’s decision tree would be valuable in evaluating game states for player optimization.