Christopher Pham
cpham32
Project 4a

Question 6:

- DummySet1: classification rate = 1.0, tree size = 3
  The tree performs perfectly on the DummySet1 with a small tree of only size 3. This indicates that there is one attribute whose values each map to a unique classification value in both the training and testing sets; in this case, attribute 5 has the exact opposite value of the classification value since both the attribute and classification are binary. Since only one attribute matters for the tree, the impact of noise from other attributes is significantly reduced.

- DummySet2: classification rate = 0.65, tree size = 11
  The tree performed poorly, only scoring 0.65. Since the classification of the data set is binary, this is not much better than random chance or an expected score of 0.5. Assuming the testing set is not significantly different than the training set, the training set needs to be improved. The score implies that either the data distribution is fairly random and each value of any attribute has roughly equal chance of mapping to either classification in which case nothing can be done, or the training set is not representative of the whole distribution and therefore more examples in the training set could improve the score. The tree only includes half of the attributes, which could indicate that the other half have even less impact on the classification.

- Connect4: classification rate = 0.760050, tree size = 41521
  The tree performed reasonably well on the Connect4 data set, scoring 0.76. The large training set, 65557 examples, was beneficial to achieving the good score. It was not higher, probably due to the nature of the game, since 8 moves is still early in the game and there are many more possible moves to affect the final classification. The tree is quite large with a size of 41521 due to the high number of attributes, 42,

- Car: classification rate = 0.946000, tree size = 408
  The tree performed excellently with a score of 0.946. The high score and tree size, which is neither very large or small, imply that the training set was very effective and there were at least a few attributes with high info gain that captured much of the variability in the classification. Since some attributes were able to account for the uncertainty, the classification was much more accurate.

Question 7:

- Connect4: The classifier helps predict the final outcome of the game based on the configuration after 8 moves. A bot could use this in addition to value iteration or q-learning to find more favorable moves that bring it closer to winning the game. Value iteration and q-learning give the bot rewards based on how certain moves will likely bring it towards the goal state, and the classifier could modify those rewards by increasing for moves that correspond to board configurations with a win classification value and decrease rewards that correspond to board configurations with a lose classification. This should enhance and expedite the decision-making process of the bot so that it makes favorable moves more often and more quickly reaches the goal.

- Cars: A data set that classifies products, like cars, by various features including physical characteristics, functionality, or brand could help a seller sort products easily. The seller could then make judgments about each product such as different pricing based on the classification or which products to stock up on since they are more likely to sell better. The classification could also help determine which products to advertise to different groups of potential customers in order to increase the effectiveness of ads and maximize return. New products or inventory could be classified in order to make the best decision on how to handle them.