

Question 6: Performance

Dummy Dataset 1

Tree size: 3. Classification rate is 1.0.

The accuracy rate is 100% and the tree size is small because attribute 5 entirely determines the label for each example. The decision tree was built accordingly, and the testing data held to the same pattern.

Dummy Dataset 2

Tree size: 11. Classification rate is 0.65.

The tree splits on 5 attributes in only 20 examples. The training data set was too small, likely making the tree rely on some attributes to make classifications when they were not the deciding attributes in the testing data.

Connect4 Dataset

Tree size: 41521. Classification rate is 0.762100.

The data set is immensely huge, as is the tree. The “low” classification rate given the size of the data set likely lies in the game, where any given combination of attributes (board positions) does not deterministically predict the winner, especially if the board is mostly blank (many possible future moves, no guarantee of optimality). The tree is large because it encodes permutations of possible board positions, of which there are very many.

Car Dataset

Tree size: 408. Classification rate is 0.945250

There are fewer attributes and relatively plenty of examples, leading to a smaller tree and higher classification rate compared to Connect 4. Because the data set covers many possible permutations of the attributes, the tree can classify testing examples very accurately.

Question 7: Applications

For the cars dataset, a similar dataset might be Kelley Blue Book. A decision tree that classifies new and used car values based on Make, Model, Year, Color, Powertrain, etc. can be extremely useful for a website to generate predictions of any future car's fair purchase price to help inform customers, or possibly even for the company to buy and sell cars for themselves and make a profit.

For the Connect 4 dataset, a decision tree based on board states might behave somewhat like a value iteration agent, where class labels (e.g. win, lose) propagate upstream and can give expected values for any given decision based on the cumulative values of the options available thereafter. Such a decision tree can help a Connect 4 bot make efficient decisions to optimize its gameplay (i.e. by choosing leaf nodes that predict victory, or normal nodes that have high expected values for winning later).