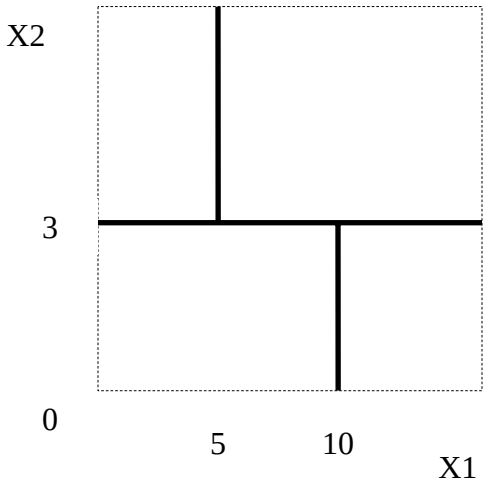
NAME:

Data Mining Techniques: Quiz 1

1. Consider two continuous attributes X1 and X2.
   a) For each of the following three partitionings, state if it could have been created by the decision tree construction algorithm we discussed in class. (6 points)
   b) If your answer is yes, draw a decision tree that corresponds to the partitioning. (24 points)

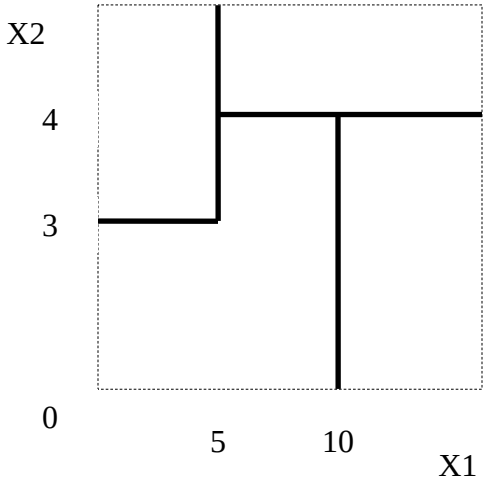


Yes.

Root split: X2 < 3

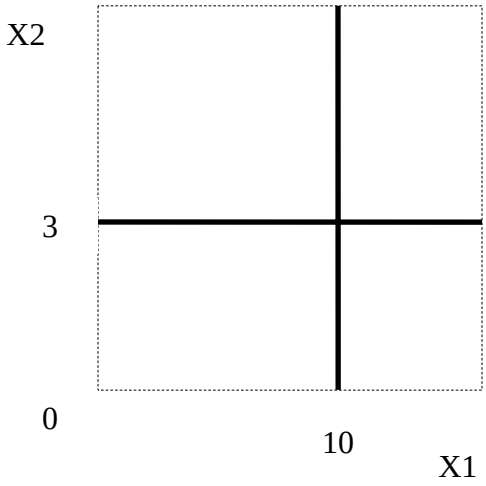
Yes-child: X1 < 10
No-child: X1 < 5

(It could also be "<=" instead of "<".)



No.

The algorithm we discussed in class only performs axis-parallel splits across the entire range of a partition.



Yes.

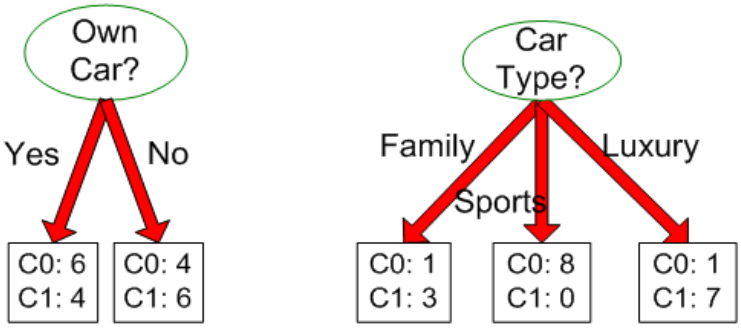
Root split: X2 < 3

Yes-child: X1 < 10
No-child: X1 < 10

(It could also be "<=" instead of "<".)

One could also split on X1 in the root and on X2 in the children.

2. Assume during tree construction you have the choice between the two split options pictured below. Which one would the algorithm we discussed in class choose? What is the justification for making this choice? (10 points)



Answer: The algorithm would select the right option. The algorithm greedily selects a split attribute that improves "purity", i.e., achieves a good separation between the two classes.

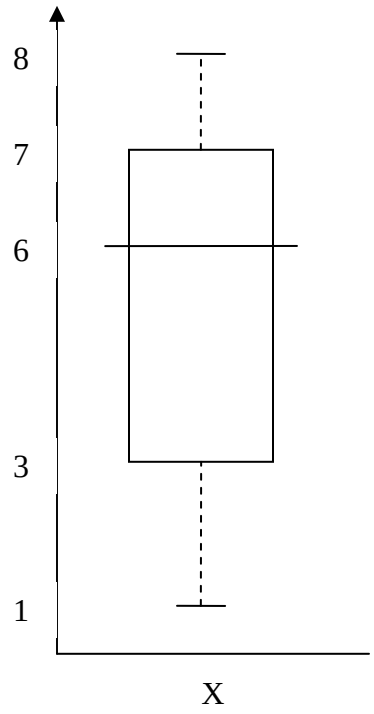
3. a) Define as precisely as possible what we mean when we say that a model *overfits*. (15 points)

Answer: A model M overfits to the training data if M performs better than another model M' on the training data, but it actually performs significantly worse than M' on the entire distribution of data tuples.

b) Why is overfitting a problem in practice? (15 points)

Answer: The model captures the idiosyncrasies of the training data, including noise, and hence might make poor predictions for other data tuples that were not used for training.

4. What does this boxplot tell you about attribute X? (20 points)



Max = 8
Q3 = 7
Median = 6
Q1 = 3
Min = 1

5. a) What does the term “training data” refer to? (5 points)

Answer: This is the set of tuples used for training a classification or prediction model.

Bonus answer: There is an implicit assumption that the training data represents the entire data distribution well.

b) What does the term “test data” refer to? (5 points)

Answer: This is the set of tuples used for evaluating the predictive performance of a model.

Bonus answer: Test tuples should not have been used for training the model. There is also an implicit assumption that the test data represents the entire data distribution well.