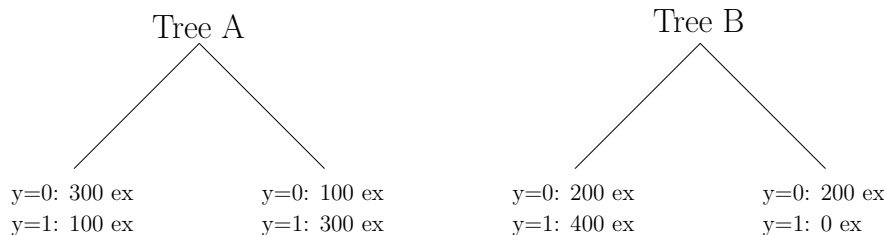


HW1: Decision trees and generalization

1 Written Exercises

Answer the following questions in 25-100 words each:

1. Consider a data set consisting of 400 data points from class 0 and 400 data points from class 1. Suppose that a tree model \mathcal{A} splits these into (300, 100) at the first leaf node and (100, 300) at the second leaf node. (Here, (n, m) denotes that n points are assigned to class 0 and m points are assigned to class 1.) Similarly, suppose that a second tree model \mathcal{B} splits them into (200, 400) and (200, 0). (See the figure below.) Evaluate the misclassification rates for the two trees: are they equal or not? Similarly, evaluate the information gain for the two trees and use these to compare the trees. Do you get different answers? Does this make sense?



2. A cousin of the decision tree is the *decision list*. In a decision list, the “left” child of any branch must be a leaf. In programming terms, a decision list has the form “if X_1 then return C_1 else if X_2 then return C_2 else if \dots ”, where the X_i s are features and the C_i s are classes. Do you imagine that entropy or the Gini index would be a good criteria for building decision lists? Why or why not? Can you think of something that might do better?
3. What purpose does the idea of “development” (or “validation”) data serve? What about cross-validation? Why is it important to do these things?
4. Why can we not just estimate hyperparameters on the training data?