

# Intermediate Exam Machine Learning, Version B

Duration of the exam: 2 hours

October 25, 2007

**You are allowed to use a calculator for this exam. It will be graded as follows: You start with 1 point, and for each of the nine subquestions you can get 1 point. Partial points may be awarded for partially correct answers. Good luck!**

1. For each of the following learning problems, please indicate whether it is a prediction, regression or classification problem. (An explanation is not required.)
  - (a) A biologist has given different amounts of food to different rats in his laboratory. He has recorded the weight of each rat after two months. Now he wants to learn how the weight of the rats depends on the amount of food they get.
  - (b) Each spring a farmer counts the number of newborn sheep. Based on his counts of the previous years he wants to estimate the number of newborn sheep in the coming year.
  - (c) A computer program tries to determine whether a newspaper article is about politics based on the number of times the article contains the following words/phrases: 'law', 'sports', 'newspaper', 'hockey', 'elections', 'human rights' and 'party'.
2. Suppose we run the LIST-THEN-ELIMINATE algorithm on the data in Table 1 with the hypothesis space  $\mathcal{H}$  that contains the four decision trees in Figure 1.
  - (a) Which trees are left in the version space after running the algorithm?
  - (b) How would the algorithm classify a new feature vector with  $A = \text{True}$ ,  $B = \text{False}$ ,  $C = \text{False}$ ?
  - (c) How would the algorithm classify a new feature vector with  $A = \text{True}$ ,  $B = \text{False}$ ,  $C = \text{True}$ ?
  - (d) Come up with a new hypothesis that is consistent with the data in Table 1, but is not contained in  $\mathcal{H}$ . Your hypothesis needs to be semantically different from all members of  $\mathcal{H}$ .
3. Suppose we first run the ID3 algorithm and then perform reduced-error pruning.
  - (a) How does reduced-error pruning change the preference bias of our algorithm (compared to running ID3 without pruning)?
  - (b) The decision to prune a node of the tree is based on the accuracy of the resulting tree on a validation set. What would go wrong if we used the train set instead of this validation set?  
Hint: See Figure 2, which is also in Mitchell.

Table 1: Boolean-valued data

$x$			$y$
$A$	$B$	$C$	
False	True	False	False
True	True	False	True
False	True	True	False

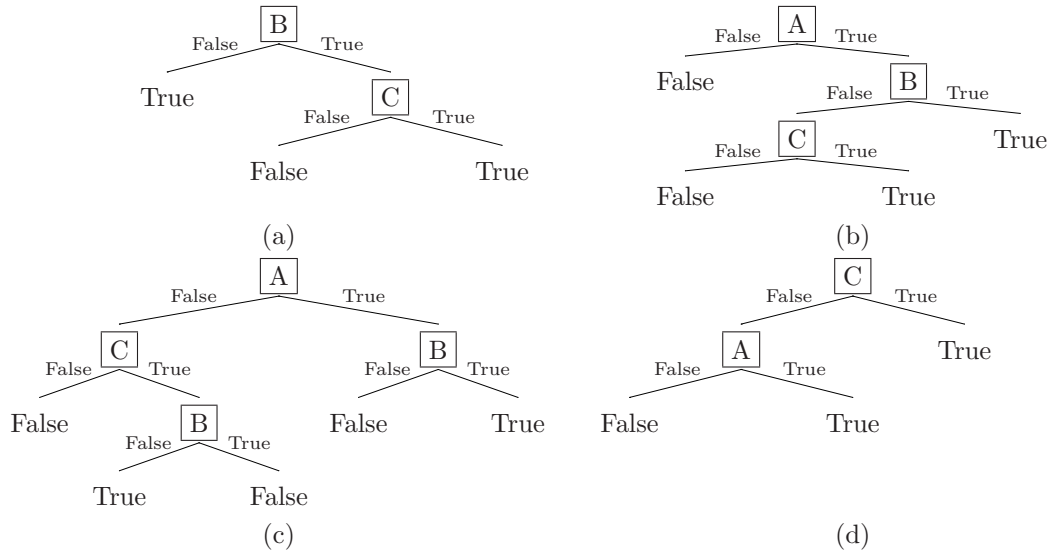


Figure 1: Four decision trees

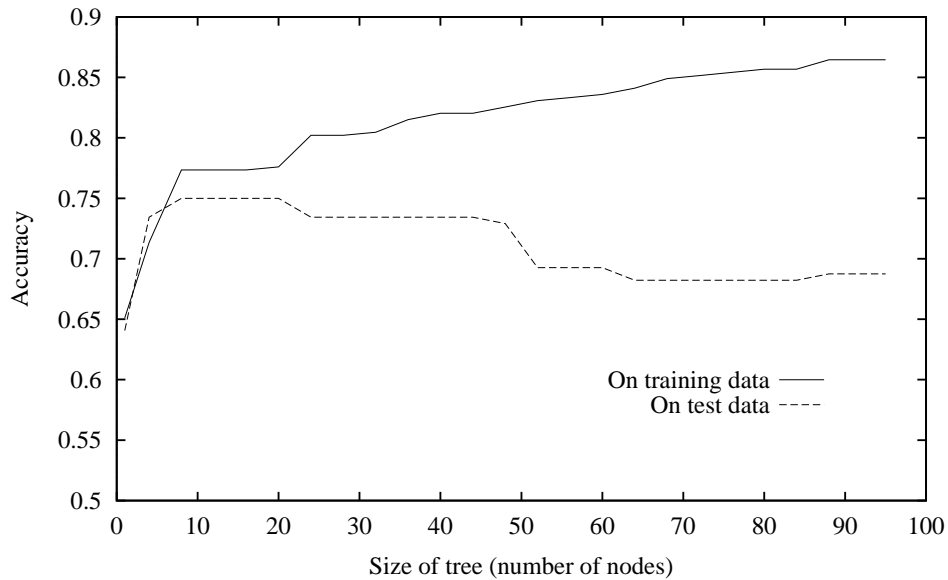


Figure 2: Accuracy of the decision tree learned by ID3 as the algorithm adds more nodes