Answers 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'.

   Answers:

   (a) regression
   (b) prediction
   (c) classification

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?

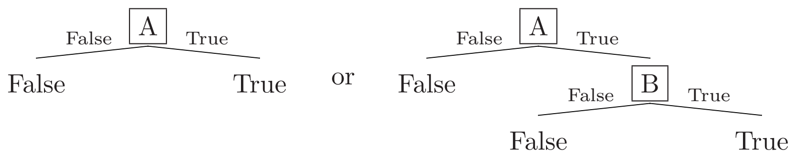
   (b) How would the algorithm classify a new feature vector with $A =$ True, $B =$ False, $C =$ False?

   (c) How would the algorithm classify a new feature vector with $A =$ True, $B =$ False, $C =$ 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 $\mathcal{H}$.

Answers:

(a) Left in the version space: (b) and (c).
(b) False
(c) Cannot classify
(d) Many different answers are possible. For example,



Using Mitchell's notation, these hypotheses might equivalently be described using a list of constraints:

$$\langle \text{True}, ?, ? \rangle \quad \text{or} \quad \langle \text{True}, \text{True}, ? \rangle,$$

and I would probably write:

$$h(\mathbf{x}) = \begin{cases} \text{True} & \text{if } A = \text{True}, \\ \text{False} & \text{otherwise}, \end{cases} \quad \text{or} \quad h(\mathbf{x}) = \begin{cases} \text{True} & \text{if } A = \text{True and } B = \text{True}, \\ \text{False} & \text{otherwise}. \end{cases}$$

All of these answers are equivalent.

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.

Answers:

(a) The algorithm will prefer even smaller trees, which are not necessarily consistent with all the training data.
(b) No pruning at all would occur, because removing nodes would decrease performance on the train set.
To see this, see Figure 2 or consider the following argument: ID3 stops adding nodes as soon as it has found a tree that is consistent with all training data. Hence removing any node from this tree will result in a tree that is not consistent with all the training data. Thus any pruning will lower the accuracy on the training data.
The purpose of pruning is to prevent overfitting. Hence when no pruning occurs, our algorithm is likely to overfit the training data.

Grading notes:

- The preferred answer is that no pruning would occur.

- The observation that it is likely that overfitting will occur is also acceptable, but only if it is correctly motivated and the word ‘overfitting’ is mentioned.
- The following answer is not correct: The tree will achieve high accuracy on the train set and **as a consequence** overfitting or worse generalisation performance will occur. This is not correct, because high accuracy on the train set does not imply bad generalisation performance and is even desirable in general (unless we have to search a very large hypothesis space to achieve it, which would cause overfitting).

Table 1: Boolean-valued data

	$\mathbf{x}$		y
A	B	C	
False	True	False	False
True	True	False	True
False	True	True	False

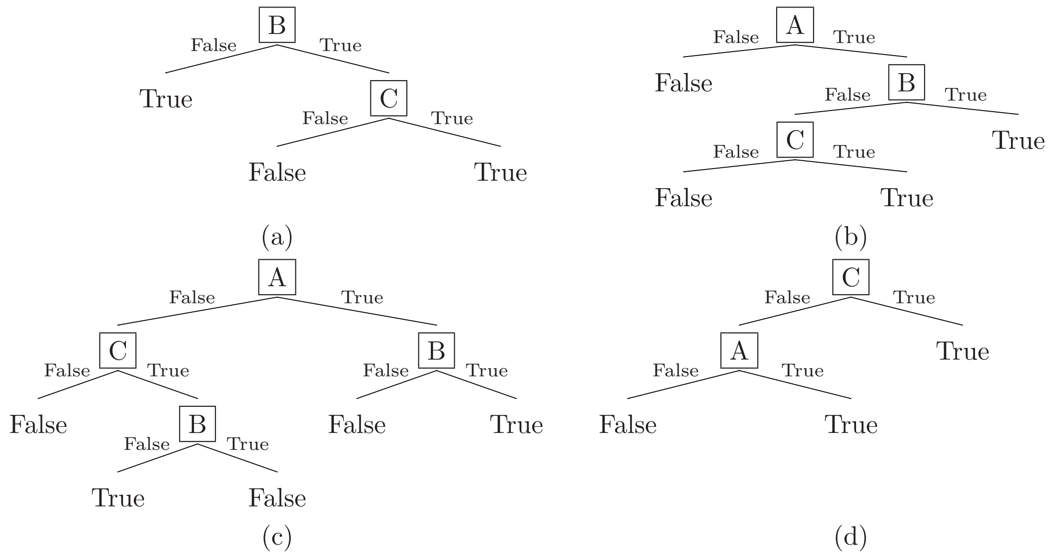


Figure 1: Four decision trees

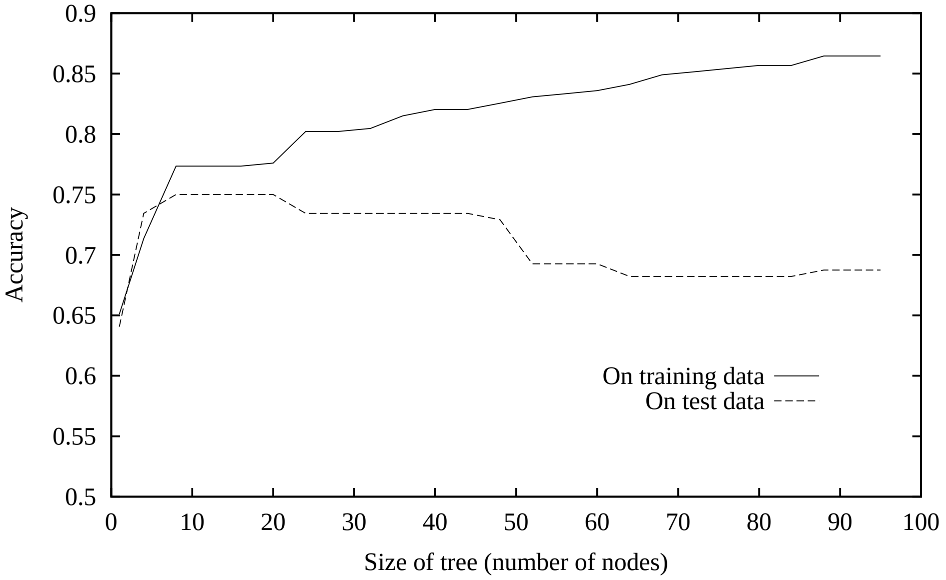


Figure 2: Accuracy of the decision tree learned by ID3 as the algorithm adds more nodes