IMPERIAL COLLEGE OF SCIENCE, TECHNOLOGY AND MEDICINE

EXAMINATIONS 2008

BEng Honours Degree in Computing Part III
MEng Honours Degree in Information Systems Engineering Part IV
MSci Honours Degree in Mathematics and Computer Science Part III
MSc in Advanced Computing
MSc in Computing Science (Specialist)
for Internal Students of the Imperial College of Science, Technology and Medicine

This paper is also taken for the relevant examinations for the Associateship of the City and Guilds of London Institute
This paper is also taken for the relevant examinations for the Associateship of the Royal College of Science

PAPER C395=I4.58

MACHINE LEARNING

Thursday 1 May 2008, 10:00
Duration: 120 minutes

Answer THREE questions

Paper contains 4 questions
Calculators required

Section A *(Use a separate answer book for this Section)*

1a Consider the following set of training examples:

Instance	Classification	a1	a2
1	+	T	T
2	+	T	T
3	-	T	F
4	+	F	F
5	-	F	T
6	-	F	T

What is the information gain of a2 relative to these training examples? Provide the equation for calculating the information gain as well as the intermediate results.

1b What is the difference between *Entropy* and *Information Gain*?

1c What is overfitting, how it can be avoided in decision tree learning, and how one can determine the correct final tree size?

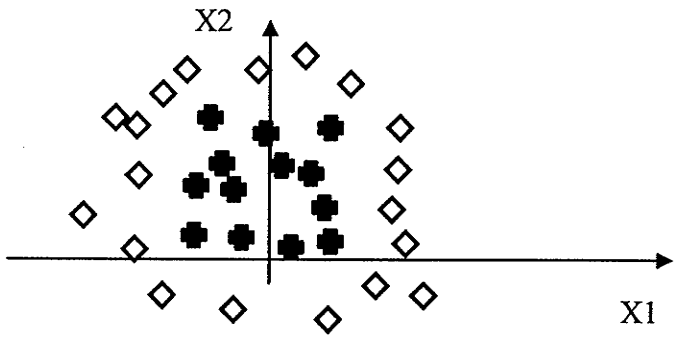
1d Consider the instance space consisting of integer points in the x, y plane, where $0 \leq x, y \leq 10$, and the set of hypotheses consisting of rectangles (i.e., being of the form $(a \leq x \leq b, c \leq y \leq d)$, where $0 \leq a, b, c, d \leq 10$).

What is the smallest number of training examples one needs to provide so that the FIND-S algorithm perfectly learns a particular target concept (e.g., $(2 \leq x \leq 4, 6 \leq y \leq 9)$)? Explain your answer in a clear manner (i.e., explain when we can say that the target concept is exactly learned in the case of the FIND-S algorithm and what is the optimal query strategy).

1e Same as Part 1d with CANDIDATE-ELIMINATION instead of FIND-S.

The five parts carry, respectively, 20%, 15%, 15%, 20%, 30% of the marks.

2 Suppose that we want to build a neural network that classifies two dimensional data (i.e., X = [x1, x2]) into two classes: diamonds and crosses. We have a set of training data that is plotted as follows:



2a Is this a supervised or an unsupervised learning problem? Explain your answer in a clear and compact manner.

2b What type of network will you choose (e.g., perceptron)? Explain your answer in a clear and compact manner.

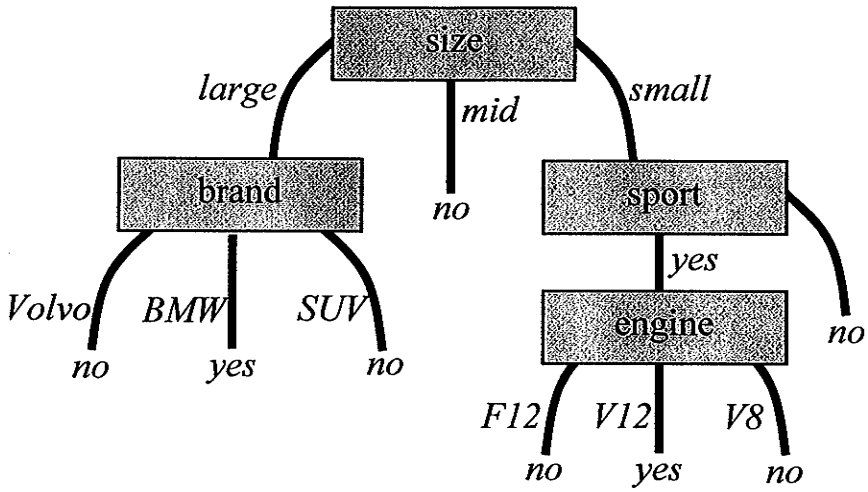
2c Draw a network that can solve this classification problem. Justify your choice of the number of nodes and the architecture. Draw the decision boundary that your network can find on the diagram.

2d Are Neural Networks an eager or a lazy learning method? What are the differences (if any) between eager and lazy learning methods? Explain your answer in a clear and compact manner.

2e What is the difference between the Best First Search and the Beam Search algorithms?

The five parts carry, respectively, 10%, 10%, 30%, 25%, 25% of the marks.

3 Suppose that we want to solve the problem of finding out what a good car is by using genetic algorithms. Suppose further that the solution to the problem can be represented by a decision tree as follows:



3a What is a suitable chromosome design for the given problem? Provide a short explanation of the solution.

3b What does "fitness function" refer to in Genetic Algorithms applications and how it can be defined? What is the suitable fitness function for the given problem? Provide a short explanation of the solution?

3c Which other Genetic Algorithm parameters need to be defined? What is the suitable definition of those parameters for the given problem? Provide a short explanation for each parameter.

3d What is the result of applying a single round of the prototypical Genetic Algorithm? Explain your answer in a clear and compact manner by providing the pseudo code of the algorithm.

The four parts carry, respectively, 20%, 20%, 30%, 30% of the marks.

END Section A *(Use a separate answer book for question 4)*