DUBLIN INSTITUTE OF TECHNOLOGY
KEVIN STREET, DUBLIN 8

BSc (Hons) in Computer Science

Stage 4

SUPPLEMENTAL EXAMINATIONS 2008

*** *SOLUTIONS* ***

ARTIFICIAL INTELLIGENCE 2

Dr. John Kelleher
Prof. B. O'Shea
Dr. I. Arana

Duration: 2 Hours

Answer Question 1 (40 marks) **and**

any 2 Other Questions (30 marks each).

*** *SOLUTIONS* ***

*** *SOLUTIONS* ***

1. (a) The logical operator $\iff$ is read *if and only if*. $P \iff Q$ is defined as being equivalent to $(P \Rightarrow Q) \land (Q \Rightarrow P)$. Based on this definition, show using truth tables the logical equivalence of $P \iff Q$ and $P \lor Q \Rightarrow P \land Q$.

    (5 marks)

P	Q	$(P \Rightarrow Q) \land (Q \Rightarrow P)$	$P \lor Q$	$P \land Q$	$(P \lor Q) \Rightarrow (P \land Q)$
T	T	T	T	T	T
T	F	F	T	F	F
F	T	F	T	F	F
F	F	T	F	F	T

   (b) Distinguish between **false negatives** and **false positives**.

    (5 marks)

> **False negative** an example can be a false negative for the hypothesis, if the hypothesis says it should be negative but in fact it is positive.
>
> **False positive** an example can be a false positive for the hypothesis, if the hypothesis says it should be positive but in fact it is negative.

   (c) In the context of machine learning, explain what is meant by **overfitting** the training data.

    (5 marks)

> Overfitting occurs when classifiers make decisions based on accidental properties of the training set that will lead to errors on the test set (or new data). As a result, whenever there is a large set of possible hypotheses, one has to be careful not to use the resulting freedom to find meaningless "regularity" in the data.

   (d) In the context of inductive learning explain what is meant by a **consistent hypothesis**.

    (5 marks)

> A hypothesis is consistent if it agrees with the true function on all examples that we have.

   (e) Describe the problems associated with measuring **classifier performance** using a single accuracy figure. Describe a more appropriate alternative.

    (20 marks)

> This is a discursive question so giving a precise answer is not appropriate. However, an answer to this question should describe how a single accuracy figure can hide a classifier's real performance. An example should be provided such as the following:
>
> - Text cases: 1000
>
> - Positive examples: 900
>
> - Negative examples: 100
>
> Assuming the classifier always classifies positively then its accuracy on the given text set would be 90% which is not an accurate reflection of the classifiers performance.
>
> The most obvious alternative would be to describe the use of **specificity**, **sensitivity** and **precision** along with a **confusion matrix**. Students should explain how a confusion matrix can be use as follows:
>
> | Classifier Results | | | |
> | --- | --- | --- | --- |
> | Class A (yes) | Class B (no) | | |
> | Correct | F_n | Class A (yes) | Expected Results |
> | F_p | Correct | Class B (no) | |
>
> And finally it is expected that the students would describe specificity, sensitivity and precision as follows:
>
> $sensitivity = \frac{t_{pos}}{pos} \; specificity = \frac{t_{neg}}{neg} \; precision = \frac{t_{pos}}{t_{pos}+f_{pos}}$

2. (a) Consider the domain of dealing 5-card poker hands from a standard deck of 52 cards, under the assumption that the dealer is fair.

   (i) Given that the number of combinations of r objects that can be selected, without regard to order and without repetition, from n distinct objects is given by the equation $\binom{n}{r} = \frac{n!}{(n-r)!r!}$, how many atomic events are there in the joint probability distribution (i.e., how many 5-card hands are there)?

   (5 marks)

   > There are $\binom{52}{5} = \frac{52!}{(47!5!)} = \frac{52 \times 51 \times 50 \times 49 \times 48}{1 \times 2 \times 3 \times 4 \times 5} = 2,598,960$ possible five-card hands

   (ii) What is the probability of each atomic event?

   (5 marks)

   > By the fair-dealing assumption, each of these is equally likely. Each hand therefore occurs with probability 1/2,598,960.

   (iii) What is the probability of being dealt a royal flush (i.e. being dealt a hand containing Ace, King, Queen, Jack and 10 all from the one suit)?

   (5 marks)

> There are four hands that are royal straight flushes (one in each suit). These events are are mutually exclusive, therefore the probability of a royal ush is just the sum of the probabilities of the atomic events, i.e., $\frac{4}{2,598,960} = \frac{1}{649,740}$.

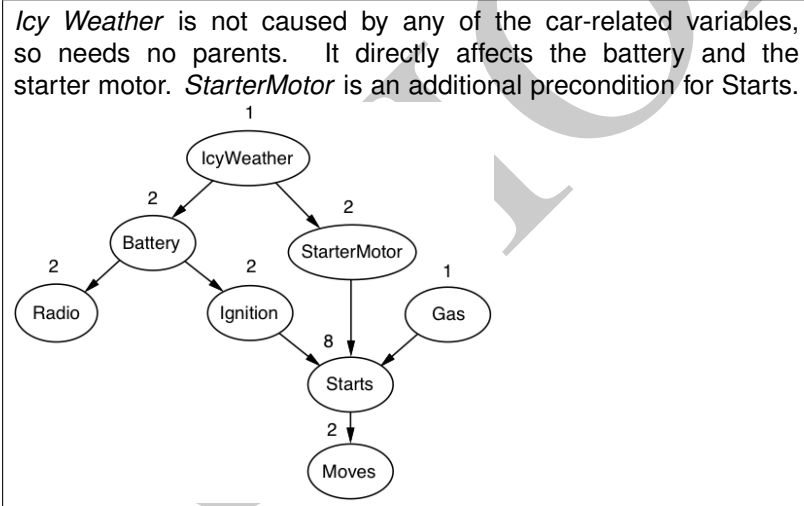(iv) What is the probability of being dealt a four of a kind (i.e. four kings, or four nines etc.)?

(5 marks)

> There are 13 possible kinds and for each, the fth card can be one of 48 possible other cards. The total probability is therefore $\frac{13 \times 48}{2,598,960} = \frac{1}{4,165}$.

(b) Consider the Bayesian network for car diagnosis shown in Figure 1.

(i) Extend the network with the Boolean variables *IcyWeather* and *StarterMotor*. *Note, cold weather normally affects the battery and start motor of a car.*

(5 marks)

> *Icy Weather* is not caused by any of the car-related variables, so needs no parents. It directly affects the battery and the starter motor. *StarterMotor* is an additional precondition for Starts.
>
>

(ii) How many independent values are contained in the joint probability?

(5 marks)

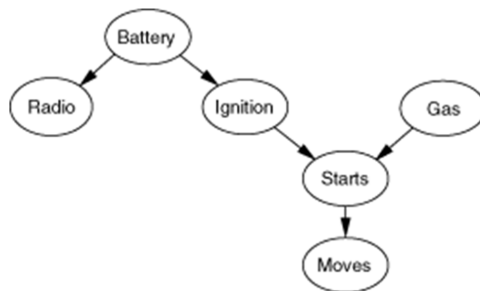> With 8 Boolean variables, the joint has $2^8 - 1 = 255$ independent entries.

Figure 1: A Bayesian network of a car's electrical system and engine. Each variable is boolean and the true value indicates that the corresponding aspect of the vehicle is in working order.

3. (a) Suppose we generate a training set from a decision tree and then apply decision-tree learning to the training set. Is it the case that the learning algorithm will eventually return the correct tree as the training set size goes to infinity? Why or why not?

(5 marks)

> The algorithm may not return the"correct" tree, but it will return a tree that is logically equivalent, assuming that the method for generating examples eventually generates all possible combinations of input attributes. This is true because any two decision tree defined on the same set of attributes that agree on all possible examples are, by denition, logically equivalent. The actual form of the tree may differ because there are many different ways to represent the same function. (For example, with two attributes A and B we can have one tree with A at the root and another with B at the root.) The root attribute of the original tree may not in fact be the one that will be chosen by the information gain heuristic when applied to the training examples.

(b) In the context of inductive logic learning, what is meant by the **extension** of a hypothesis?

(5 marks)

> Each hypothesis predicts that a certain set of examples, namely those that satisfy the hypotheses definition, are examples that satisfy the goal predicate. This set of examples is called the extension of the hypothesis. For example, assuming a standard interpretation, the extension of the predicate $digit(X)$ is $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 0\}$

(c) Distinguish between the **generalisation** and **specialization** of a logical predicate.

(10 marks)

> The generalisation of a logical predicate results in the broadening of the extension of the predicate. Some of the operations that can be used to generalise a logical predicate include: converting conjunctions in the predicate to disjunctions or dropping conditions in the predicate. You would generalise a predicate in response to recognising a false negative. The specialisation of a logical predicate results in the narrowing of the extension of the predicate. Some of the operations that can be used to generalise a logical predicate include: converting disjunctions in the predicate to conjunctions or adding conditions in the predicate. You would specialise a predicate in response to recognising a false positive.

(d) For some data sets it is possible to devise multiple hypotheses that are consistent with the data. Describe a heuristic for choosing among multiple consistent hypotheses and explain why your heuristic is reasonable.

(10 marks)

> One answer is to use Occams razor (sometimes called Ockhams razor) : prefer the hypothesis that maximizes a combination of simplicity and consistency with the data. This makes sense, because hypotheses that are no simpler than the data themselves are failing to extract any pattern from the data. Defining simplicity is not easy but it seems reasonable to say that a degree-1 polynomial is simpler than a degree-12 polynomial.

4. (a) What does it mean if two classes C_1 and C_2 are described as **linearly separable**.
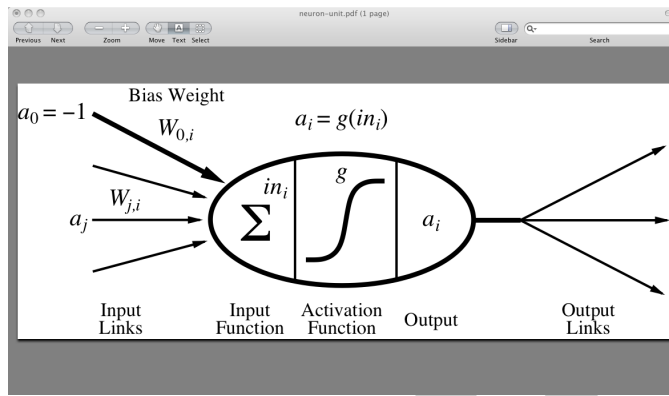
(5 marks)

> This means that for each class C_i there exists a hyperplane H_i such that on its positive side lie all $x \in C_i$ and on its negative side lie all $x \in C_j, j \neq i$

(b) Describe the processing stages of a McCulloch-Pits "unit".

(10 marks)

The processing stages of a unit are:

(i) Each unit i first compute a weighted sum of its inputs: $in_i \leftarrow \sum_j W_{j,i} a_j$

(ii) Then it applies an **activation function** g to this sum to derive the output (activation) a_i: $a_i \leftarrow g(in_i) = g\left(\sum_j W_{j,i} a_j\right)$

(c) Construct by hand a neural network that computes the XOR function of two inputs. Make sure to specify what sort of units you are using.

(15 marks)

XOR (in fact any Boolean function) is easiest to construct using step-function units. Because XOR is not linearly separable, we will need a hidden layer. It turns out that just one hidden node sufces. To design the network, we can think of the XOR function as OR with the AND case (both inputs on) ruled out. Thus the hidden layer computes AND, while the output layer computes OR but weights the output of the hidden node negatively. The image below illustrates the network.