香 港 中 文 大 學
The Chinese University of Hong Kong

Course Examination 1st Term, 2010 - 2011

Course Code & Title : CSCI3230 Fundamentals of Artificial Intelligence

Time allowed : 2 hours 0 minutes

Student I.D. No. : Seat No. :

Answer Four out of Five questions
Total: 100 marks

1.

a) What is the crucial difference between the following two first-order sentences where R is a binary predicate? [3 marks]

i. $\forall x \exists y\, R(x, y)$.

ii. $\exists x \forall y\, R(x, y)$.

b) Express each of the following statements (i to v) as First Order Logic sentences using the symbols below (together with $\forall, \exists, \neg, \wedge, \vee, \rightarrow$): [5 marks]

Symbol	Meaning
Student(X)	X is a student
Diligent(X)	X is diligent
Lazy(X)	X is lazy
DoRevision(X,Y)	X will do revision for course Y
Tough(Y)	Y is a tough course
Pass(X,Y)	X passes the course Y

i. Every student is either diligent or lazy.
ii. Diligent students will do revision for all courses.
iii. Lazy students cannot pass any tough course.
iv. *CSCI3230* is a tough course, and *Peter* is a student.
v. If *Peter* does not do revision for *CSCI3230*, then he cannot pass it.

c) Convert your answers for b) into **Conjunctive Normal Form (CNF)**. [6 marks]

d) Prove statement v from statements i to iv using *any one* of the **Forward Chaining**, **Backward Chaining** and **Resolution** approaches. [6 marks]

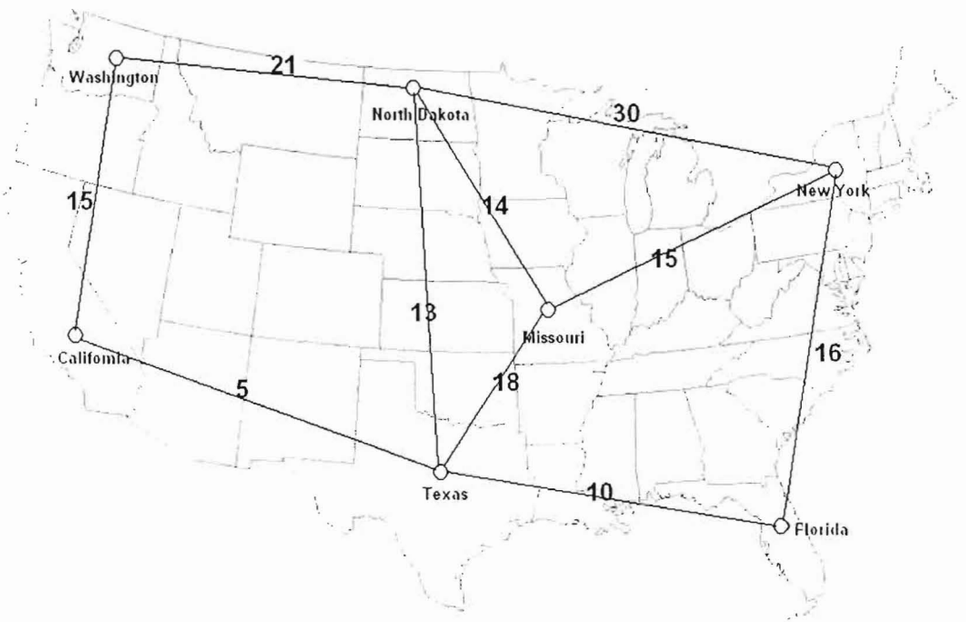
e) What are **soundness** and **completeness** for a set of inference rules? What are their importances? [5 marks]

2.

a) Name a single point stochastic search algorithm and describe it briefly. [3 marks]

b) The distances between cities and the Heuristic Distances (HD) to New York are shown in the U.S.A. roadmap and the table below respectively. Use **both** of the following two algorithms to search for a shortest path from **Washington** to **New York**. Show all your steps and state the paths, costs and sequence (step order numbers) clearly with a path tree.
(In each step, you do NOT need to consider the city you have just visited)

i. Uniform-cost search with path costs [6 marks]
ii. A* search (using the given Heuristic Distances table below) [8 marks]



Heuristic Estimated distances to New York				
New York	0		North Dakota	28
Washington	20		California	40
Missouri	34		Florida	45
Texas	43			

c) Do the two algorithms used for part b) give the same shortest path, and are the given Heuristic Estimated distances admissible? Explain your answer.
[4 marks]

d) Name the heuristic searching technique used in the Multi-Layer Feed-Forward Network (MLN) back-propagation updating rule? Why is it used in MLN? [4 marks]

3.

a) Consider the following game: there are N coins, labeled 0 to N-1, some have heads facing up, and some have tails facing up. If you choose a coin to flip, its immediate left and right neighbors (note: N-1 is a neighbor of 0) will be flipped (i.e. head → tail, tail → head), but not the coin itself. E.g. choosing coin 2 will flip the coins 1 and 3; choosing coin 0 will flip the coins 1 and N-1. The objective of the game is to successively flip the coins so that all of the coins have heads facing up. Formulate the game as a search problem (state space, state transitions and goal state(s)).

[8 marks]

b) Use *any heuristic* you like to solve the following instance, where $N = 7$ and initially the coins are as follows. List which coins you successively flipped and the state transitions. [3 marks]

0	1	2	3	4	5	6
Tail	Tail	Tail	Head	Tail	Tail	Tail

c) In real problems, the search spaces are usually very large. Suggest **TWO** methods to handle the huge spaces. [4 marks]

d) In game playing (e.g. Chess, Go, Checker), even using MiniMax with Alpha-Beta pruning, it is not guaranteed that the computer can win (at least not lose). Explain why this is so. [4 marks]

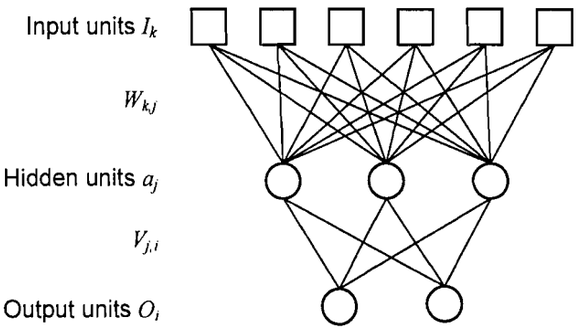
e) What is **overfitting** in supervised learning? What is **Ockham's Razor**? How does it relate to overfitting? How can we detect and prevent overfitting? [6 marks]

4.

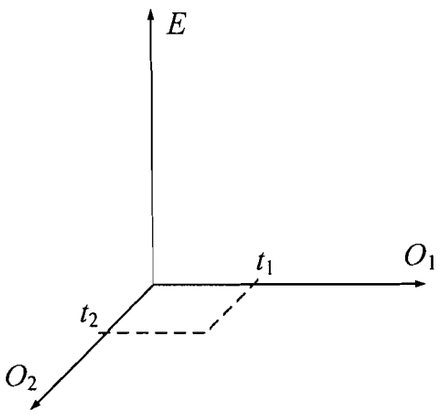
a) Why is single layer perceptron unable to solve **non-linearly separable** problems? Give an example of **non-linearly separable** problem. [3 marks]

b) For the sigmoid function $f(z)=\frac{1}{1+e^{-z}}$, find its derivative $f'(z)$. [3 marks]

c) Derive the **TWO** updating rules ($\frac{\partial E}{\partial V_{j,i}}$ and $\frac{\partial E}{\partial W_{k,j}}$) of weights and biases in back-propagation algorithm for Multi-Layer Feed-Forward Network (MLN). The activation function in the output units is the above sigmoid function, i.e., $O_i = f(u_i)$, where $u_i = \sum_j a_j \times V_{j,i} + B_{O_i}$. The activation function in the hidden units is the above sigmoid function, i.e., $a_j = f(s_j)$, where $s_j = \sum_k I_k \times W_{k,j} + B_{a_j}$. The error function is the sum of squared error, i.e., $E = \frac{1}{2}\sum_i (O_i - T_i)^2$ where T_i is the target value. The notations and the diagram of the MLN are shown below: [12 marks]



d) Why $E = \frac{1}{2}\sum_i (O_i - T_i)^2$ is suitable to be used as an error function in back-propagation algorithm of MLN? Explain briefly. (Hint: Sketch the E where t_1 and t_2 are the target values of O_1 and O_2 respectively). [7 marks]

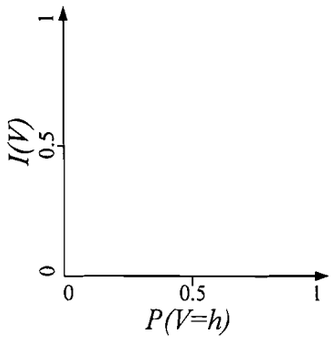


5.

a) For a discrete random variable V with n possible values v_i's, each with probability P_i, the Information Content (I) is given by:

$I(V) = \sum_{i=1}^{n} -P_i \log_2 P_i$ with the convention that $0 \times \log_2 0 = 0$

For a binary random variable (e.g. for flipping a coin, $n = 2$), $P(V = h)$ denotes the probability of flipping a "head". **Sketch** the graph of $I(V)$ against $P(V = h)$ on your answer book, and describe and explain it briefly (such as the maxima, minima and the convexity). [5 marks]



b) For a K class classification, an attribute with N values $A \in \{a_1, a_2, ..., a_N\}$, and $v_{i,k}$ is the number of class-k records having value a_i in that attribute. Write down the equation in computing the Information Gain (IG) of the attribute. [3 marks]

c) In constructing a Decision Tree, the attribute having the highest IG is chosen to be the splitting attribute. Explain why. [2 marks]

d) Note that the decision tree constructed by the method in part c) may NOT be optimal, why? Suggest a method to improve the construction method. [4 marks]

e) In binary classification, the confusion matrix consists of four elements which are True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN). What does each mean? [4 marks]

f) There are a number of performance measures derived from TP, FP, TN and FN:

$Precision = \frac{TP}{TP + FP}$, $Recall = \frac{TP}{TP + FN}$, $F\_measure = \frac{1}{\left(\frac{1}{Recall} + \frac{1}{Precision}\right)/2}$

Interpret each of them. [4 marks]

g) In a medical classification, e.g. predicting whether a patient is suffering from cancer or not, which performance measure(s) in part f) is more important. Explain why. [3 marks]

-- END --