Student ID: .....

TE WHARE WĀNANGA O TE ŪPOKO O TE IKA A MĀUI

VICTORIA
UNIVERSITY OF WELLINGTON

EXAMINATIONS – 2017

TRIMESTER ONE

COMP 307

INTRODUCTION TO ARTIFICIAL INTELLIGENCE

Time Allowed: TWO HOURS

CLOSED BOOK

Permitted materials: Only silent non-programmable calculators or silent programmable calculators with their memories cleared are permitted in this examination.
Non-electronic foreign language translation dictionaries may be used.

Instructions: There are a total of 120 marks on this exam.
Attempt all questions.
The appendix on the last sheet can be torn off for reference for questions 2-6.

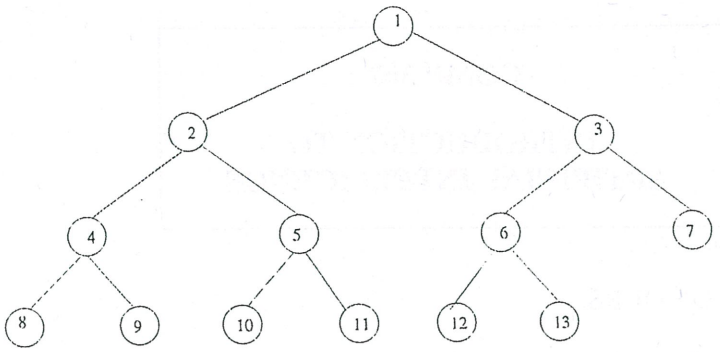
Questions

1. Search [15]
2. Machine Learning Basics [20]
3. Neural Networks [10]
4. Evolutionary Computation and Learning [15]
5. Reasoning under Uncertainty [10]
6. Bayesian Networks [20]
7. Planning and Scheduling [20]
8. Other Topics [10]

Question 1. Search

[15 marks]

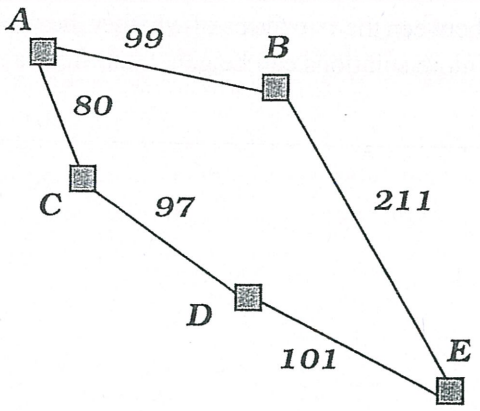
Based on the figure below, answer questions (a) and (b).



(a) [1 mark] Assuming that you are using *breadth first* search, state the search order/path using the numbers in the nodes.

(b) [4 marks] Assuming that you are using *iterative deepening* search, state the search order/path using the numbers in the nodes.

(c) [5 marks] The figure below represents part of a city map. Assume that the numbers in the figure are the costs between two cities (nodes). When *uniform cost search* is used and the problem is to travel from city A to city E, provide the search path and the final solution.



(d) [2 marks] Assume that $h(n)$ is the estimated cost of the cheapest path from Node n to the goal state, and $g(n)$ is the cost from the start node to Node n. State the difference between the *greedy (best first) search* and the *A* search* techniques in terms of the evaluation function.

(e) [3 marks] *Beam search* and *gradient descent search* and are two heuristic search methods.

(i) State a machine learning algorithm/technique algorithm that uses *beam search*.

(ii) State a machine learning algorithm/technique that uses *gradient descent search*.

(iii) Briefly describe the differences between them in terms of whether they are local or global search techniques, and whether one or more solutions can be generated from a single experiment run.

Student ID: .....

SPARE PAGE FOR EXTRA ANSWERS

Cross out rough working that you do not want marked.
Specify the question number for work that you do want marked.

Question 2. Machine Learning Basics [20 marks]

(a) [3 marks] There are different types of learning, e.g. supervised learning, unsupervised learning, and reinforcement learning.

(i) State the main difference between *supervised* learning and *unsupervised* learning.

(ii) State a machine learning task for each of the two types of learning in part(i).

(b) [5 marks] In supervised machine learning systems, a dataset is typically separated into a *training set*, a *test set*, and a *validation set*.

(i) Briefly describe the role of each of them.

(ii) *k-fold cross validation* is a widely used method in classification. State the difference between *k-fold cross validation* and a *validation set*.

(c) [4 marks] *K-Nearest Neighbour* and *K-means* are two methods that are widely used in machine learning.

(i) For each method, state whether it is for classification, clustering, or association rules.

(ii) For *K-Nearest Neighbour*, state whether it is typically applied to numeric data or categorical data.

(iii) If you run the *K-means* method multiple times, does the method always produce the same clusters?

(d) [2 marks] The XOR problem is a binary classification problem where the target output value is 1 when the two input variables have different Boolean values, and the target output value is 0 when the two input values are the same.

If a perceptron is used to solve the XOR problem:

(i) State whether the XOR problem can be successfully solved by the perceptron.

(ii) State the major limitation of the perceptron learning algorithm.

(e) [6 marks] Consider the following dataset describing 10 pizzas from a pizza shop, of which 5 are popular with customers, and 5 are not. They are described by three attributes: whether they have mushroom or not, whether they are vegetarian or have meat, and whether they are in small, medium or large size.

Instance	Mushroom	Vegetarian	Size	Class
1	yes	meat	medium	popular
2	yes	meat	large	popular
3	yes	meat	medium	popular
4	yes	vegetarian	large	popular
5	no	meat	small	popular
6	no	meat	large	unpopular
7	yes	vegetarian	small	unpopular
8	no	vegetarian	small	unpopular
9	no	meat	medium	unpopular
10	no	vegetarian	large	unpopular

The pizza shop wants to build a *decision tree* for classifying pizzas to *popular* or *unpopular*. Which attribute should be chosen for the *root* of the decision tree if they use the weighted impurity function $P(popular) * P(unpopular)$? Show your working.

Student ID: .....

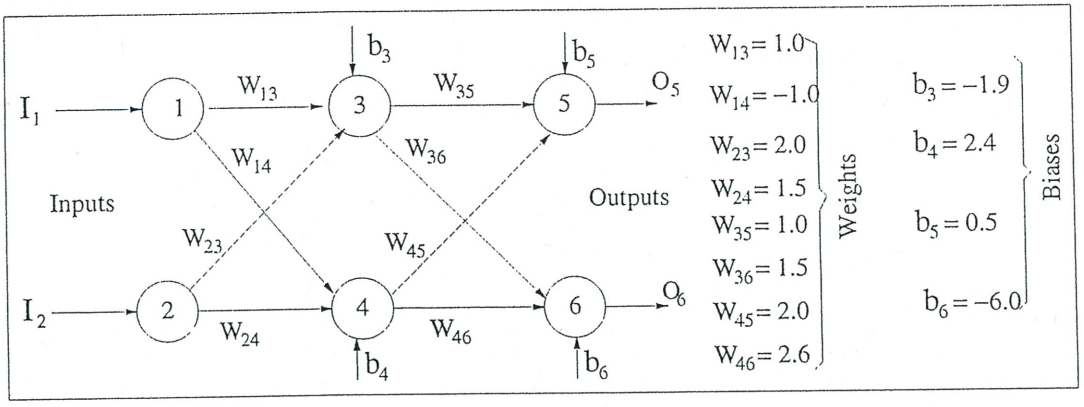
SPARE PAGE FOR EXTRA ANSWERS

Cross out rough working that you do not want marked.
Specify the question number for work that you do want marked.

Question 3. Neural Networks

[10 marks]

(a) [6 marks] Consider the following feed forward neural network that we discussed during the tutorials and uses the sigmoid/logistic transfer function (see Appendix B),



(i) What will the output of node 5 be (O_5) for the input vector (0.0, 0.0)?

(ii) What will the new value of weight W_{35} be after one epoch of training using the back propagation algorithm? Assume that the training set consists of only a single instance with an input vector (0.0, 0.0) representing input feature values for I_1 and I_2, and an output vector (0.0, 0.0) representing the target output values for nodes O_5 and O_6, and that the learning rate η is 0.15.

Show your working.

(b) [4 marks] James Browne has developed a classifier for distinguishing the quality of different wines into four classes: *excellent, good, fair*, and *bad*. He extracted 6 features from the given 400 instances, used the standard multilayer feed forward neural network, and applied the back propagation algorithm to train his network for classification. Among the 400 instances (100 for each of the four classes), he chose 80 (20 for each class) for network training and the other 320 for testing. The network architecture he used is 6-50-4 (6 input nodes, 50 hidden nodes and 4 output nodes). After training for 30,000 epochs, the network classifier obtained 98.5% accuracy on the training set, but only achieved 60% accuracy on the test set. Suggest four good ways to James for improving the (test) performance.

SPARE PAGE FOR EXTRA ANSWERS

Cross out rough working that you do not want marked.
Specify the question number for work that you do want marked.

Question 4. Evolutionary Computation and Learning [15 marks]

(a) [3 marks] Genetic algorithms use *genetic operators* to update the population from generation to generation. Three key genetic operators are *crossover*, *mutation*, and *elitism*. Briefly describe each operator.

(b) [6 marks] Briefly describe the general evolutionary process in genetic algorithms. Draw a figure if necessary.

(c) [6 marks] Genetic Programming (GP) is a good method for symbolic regression, which evolves a mathematical function to model the relationship between the output variable and the input variable(s) from a (training) set of instances. Suppose your task is to use GP to evolve a mathematical model to map a single input variable x to the single output variable y, given a set of 9 training datapoints shown in the following table.

x	-1.2	-0.9	-0.6	-0.3	0	0.3	0.6	0.9	1.2
y	-6.748	-5.119	-3.976	-3.157	-2.5	-1.843	-1.024	0.119	1.748

(i) Choose a good terminal set.

(ii) Choose a good function set.

(iii) Choose a good fitness function.

(iv) State two main differences of GP for symbolic regression over traditional statistical regression.

SPARE PAGE FOR EXTRA ANSWERS

Cross out rough working that you do not want marked.
Specify the question number for work that you do want marked.

Question 5. Reasoning under Uncertainty [10 marks]

Use the probability rules discussed in the lectures to answer questions (a) to (e). Show your working.

Reminder: the $\neg$ symbol stands for "not", the $\perp\!\!\!\perp$ symbol stands for "independence", and the vertical line stands for "given". If no value is given to the variable, e.g. just P(A), you can read it as P(A=true).

[Note: Regarding the calculation, you should show your *working* of the calculation in the form like $P(X = 0|Y = 1) = \frac{P(X=0,Y=1)}{P(Y=1)}$, to demonstrate that you *know how to calculate* them. It does not matter whether you compute the final numbers here – you can write your final answers in expression form like $\frac{0.25+0.18}{0.69}$.]

(a) [1 mark] Consider a *Boolean* variable A, if we know $P(A) = 0.45$, what is $P(\neg A)$?

(b) [2 marks] Consider two *Boolean* variables A and B, if we know $P(B|A) = 0.25$, which of the following do we also know and what is its value ?

- $P(B|\neg A)$
- $P(\neg B|A)$
- $P(\neg B|\neg A)$

(c) [1 mark] Consider two *Boolean* variables A and B, $P(B) = 0.7$, and $P(A|B) = 0.35$. If $A \perp\!\!\!\perp B$, what is the probability of $P(A)$?

(d) [3 marks] Consider three Boolean variables A, B, and C:

(i) If $A \perp\!\!\!\perp B|C$, $P(A) = 0.5$, $P(B) = 0.6$, $P(C) = 0.3$, $P(A|C) = 0.2$, and $P(B|C) = 0.4$, calculate $P(A, B|C)$;

(ii) Does $A \perp\!\!\!\perp B$ imply that $A \perp\!\!\!\perp B|C$?

(iii) Does $A \perp\!\!\!\perp B|C$ imply that $A \perp\!\!\!\perp B$?

(e) [3 marks] The table below lists the joint probabilities of three *binary* random variables X_1, X_2, and X_3.

X_1	X_2	X_3	$P(X_1, X_2, X_3)$
0	0	0	0.06
1	0	0	0
0	1	0	0.15
1	1	0	0.24
0	0	1	0.05
1	0	1	0.25
0	1	1	0.2
1	1	1	0.05

(i) What is the probability $P(X_1 = 0, X_3 = 0)$?

(ii) What is the probability of $P(X_2 = 1|X_1 = 0, X_3 = 0)$?

SPARE PAGE FOR EXTRA ANSWERS

Cross out rough working that you do not want marked.
Specify the question number for work that you do want marked.

Question 6. Bayesian Networks

[20 marks]

(a) [3 marks] This question concerns the Naive Bayes classification method we discussed in the lectures:

(i) State the assumption of the Naive Bayes learning method.

(ii) Compared with Neural Networks, Naive Bayes is a relatively fast learning algorithm. Briefly state the main reason.

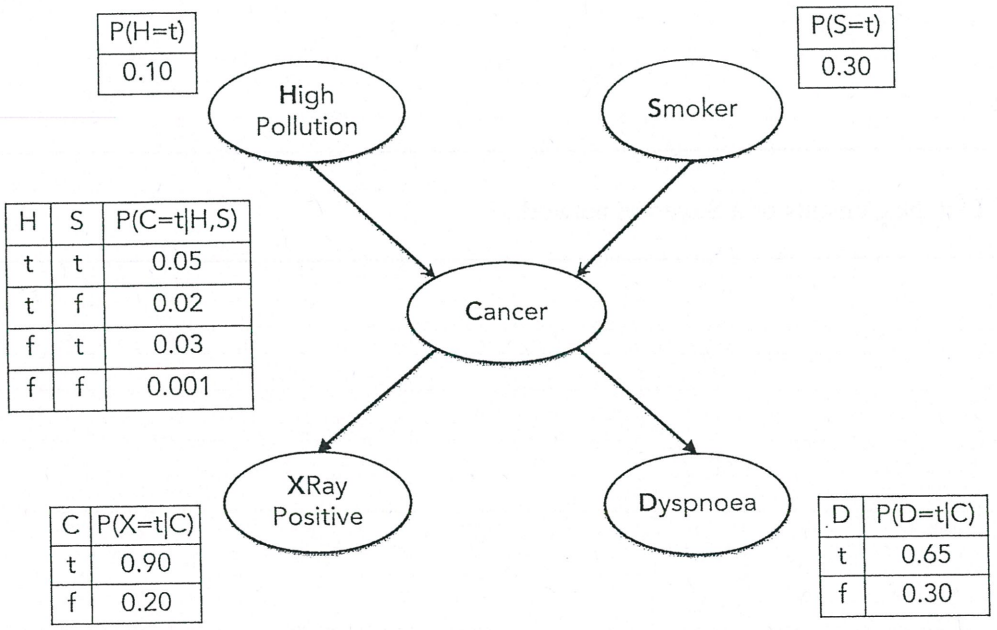
(b) [2 marks] List the elements of a Bayesian network.

The following Bayesian network represents two causes and two effects related to lung cancer.

- Being a smoker can cause lung cancer;
- Having been exposed to high air pollution can cause lung cancer;
- Lung cancer can cause dyspnoea;
- Lung cancer can cause the XRay result being positive.

Each variable takes the value *true (t)* or *false (f)*. We will abbreviate the five variable names using the bold letters shown in the figure: **H**, **S**, **C**, **X**, and **D**. The probabilities shown are all for the "is true" outcome, e.g. read $P(S = t) = 0.30$ as the probability that the variable Smoker takes the value t is 0.30. Answer questions (c) to (g).

[Note: Regarding the calculation, you should show your *working* of the calculation in the form like $P(X = 0|Y = 1) = \frac{P(X=0,Y=1)}{P(Y=1)}$, to demonstrate that you *know how to calculate* them. It does not matter whether you compute the final numbers here – you can write your final answers in expression form like $\frac{0.25+0.18}{0.69}$.]



P(H=t)
0.10

P(S=t)
0.30

H	S	P(C=t|H,S)
t	t	0.05
t	f	0.02
f	t	0.03
f	f	0.001

C	P(X=t|C)
t	0.90
f	0.20

D	P(D=t|C)
t	0.65
f	0.30

(c) [4 marks] Given the above Bayesian network, answer *true* or *false* for the following questions (note you do not need to consider the specific values in the tables to answer this question):

(i) $H \perp\!\!\!\perp D|C$

(ii) $H \perp\!\!\!\perp S|C$

(iii) $P(X|H,S,C) = P(X|C)$

(iv) $X \perp\!\!\!\perp D|C$

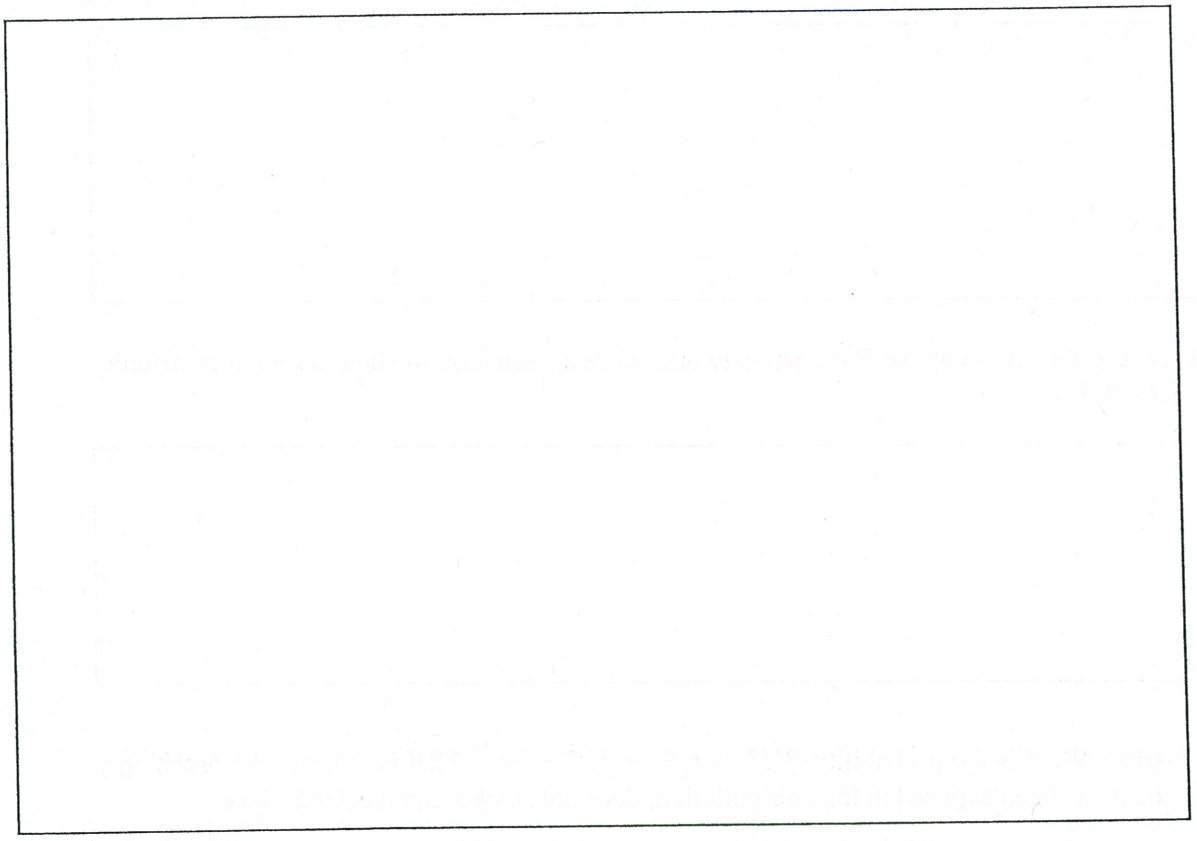
(d) [2 marks] Given the above Bayesian network, write an equation to show the joint probability $P(H,S,C,X,D)$.

(e) [2 marks] What is the probability $P(H = t, S = f, C = t)$? That is, what is the probability that a patient has been exposed in high air pollution, does not smoke, and has lung cancer ?

(f) [4 marks] As disccused in our lectures, one way to build a Bayesian network is to first choose a set of relevant variables that describe the domain, choose an order for the variables, and then each variable is added to the network in order. When adding a variable as a node, we need to add arcs to this node from *a minimal set of parents/nodes* in the existing network.

(i) Why it is important to add arcs from *a minimal set of parents* ?

(ii) If the given order is (X, D, C, H, S), draw the structure (nodes and arcs) of the Bayesian network that would be built.



(g) [3 marks] This question concerns *inference* in Bayesian networks.

Assume that the XRay result of a patient is positive, this patient has dyspnoea, and that you want to know the probability that the patient is a smoker.

(i) What are the hidden variables in the inference?

(ii) Why will a large number of hidden variables increase the computational cost? What can you do to avoid high computational cost?

Question 7. Planning and Scheduling [20 marks]

(a) [5 marks] In the lectures, we have discussed two common and related AI Topics called *planning* and *scheduling*. State the definitions of **planning** and **scheduling**, and list three applications for each.

(b) [2 marks] The STRIPS (STanford Research Institute Problem Solver) and ADL (Action Description Language) planning algorithms (planners) have been discussed during the lectures. List **four** differences between STRIPS and ADL.

(c) [1 mark] State the major difference between **progression** state-space search and **regression** state-space search in terms of search directions.

(d) [2 marks] As discussed during the lectures, a state in STRIPS is represented by a *conjunction* of *positive*, *ground*, and *function-free* literals. Which of the following items (capital words are instantiations and lowercase words are variables) are valid STRIPS state representations?

- $Go(there)$
- $Buy(MUFFINS)$
- $At(P1, WLG) \wedge Airport(WLG)$
- $Clear(A) \vee At(B, C)$

(e) [6 marks] The "Air cargo transport" problem can be represented in STRIPS as follows:

$Action(Fly(p, from, to)$,

PRECOND: $Plane(p) \wedge At(p, from) \wedge Airport(from) \wedge Airport(to)$

EFFECT: $\neg At(p, from) \wedge At(p, to))$

$Action(Load(c, p, a)$,

PRECOND: $Plane(p) \wedge Cargo(c) \wedge Airport(a) \wedge At(p, a) \wedge At(c, a)$

EFFECT: $\neg At(c, a) \wedge In(c, p))$

$Action(Unload(c, p, a)$,

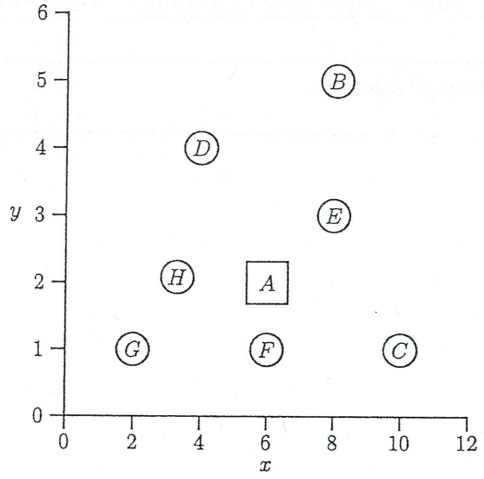
PRECOND: $Plane(p) \wedge Cargo(c) \wedge Airport(a) \wedge At(p, a) \wedge In(c, p)$

EFFECT: $\neg In(c, p) \wedge At(c, a))$

(i) Suppose the current state is $Plane(A320) \wedge Cargo(Books) \wedge Airport(WLG) \wedge Airport(CHC) \wedge At(A320, CHC) \wedge In(Books, A320)$. List all the applicable actions.

(ii) For each applicable action (based on Part (i)), give the resultant state.

(f) [4 marks] The graph below describes a vehicle routing problem. In total, there are 8 nodes including the depot (A). Apart from the depot, each node has a demand of 1. The depot node (A) has a demand of 0.



Assuming the capacity is 4, state whether each of the following solutions is feasible or not. If the solution is infeasible, justify why.

(i) $R_1 = (A, C, F, H)$, $R_2 = (A, B, D, E, G, A)$

(ii) $R_1 = (A, H, G, D, F, A)$, $R_2 = (A, B, C, A)$

(iii) $R_1 = (A, C, F, A)$, $R_2 = (A, G, B, E, H, D, A)$

(iv) $R_1 = (A, C, H, F, D, A)$, $R_2 = (A, B, G, E, A)$

Question 8. Other Topics

[10 marks]

At the final stage of our lectures, we provided a very brief discussion of *knowledge based systems, natural language processing, support vector machines, data and web mining, big data*, and *deep learning*.

(a) [3 marks] State three characteristics of *big data*.

(b) [2 marks] List two areas of *knowledge based systems* that are still widely used these days.

(c) [2 marks] List two commonly used architectures and two real-world applications of *deep learning*.

(d) [2 marks] Define the term *knowledge discovery in databases (KDD)* in context of *data mining*.

(e) [1 mark] List two real-world applications of *natural language processing*.

* * * * * * * * * * * * * * *

SPARE PAGE FOR EXTRA ANSWERS

Cross out rough working that you do not want marked.
Specify the question number for work that you do want marked.

Student ID: .....

SPARE PAGE FOR EXTRA ANSWERS

Cross out rough working that you do not want marked.
Specify the question number for work that you do want marked.

Appendix for COMP307 exam

(You may tear off this page if you wish.)

A Some Formulae You Might Find Useful

$$p(C|D) = \frac{p(D|C)p(C)}{p(D)} \tag{1}$$

$$f(x_i) = \frac{1}{1+e^{-x_i}} \tag{2}$$

$$O_i = f(I_i) = f(\sum_k w_{k\to i} \cdot o_k + b_i) \tag{3}$$

$$\Delta w_{i\to j} = \eta o_i o_j (1-o_j)\beta_j \tag{4}$$

$$\beta_j = \sum_k w_{j\to k} o_k (1-o_k)\beta_k \tag{5}$$

$$\beta_j = d_j - o_j \tag{6}$$

B Sigmoid/Logistic Function