Student ID: .....

TE WHARE WĀNANGA O TE ŪPOKO O TE IKA A MĀUI
VICTORIA
UNIVERSITY OF WELLINGTON

EXAMINATIONS – 2013

TRIMESTER ONE

COMP 307

********** **WITH**

SOLUTIONS **********

INTRODUCTION TO
ARTIFICIAL INTELLIGENCE

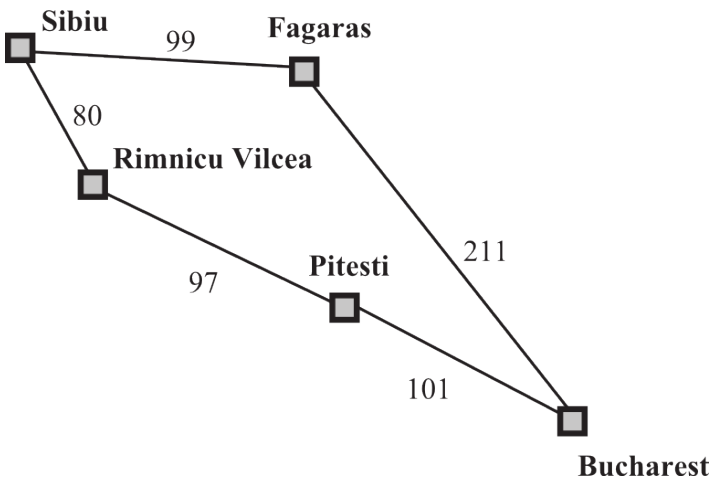
Time Allowed: THREE HOURS

Instructions: Closed Book.
There are a total of 180 marks on this exam.
Attempt all questions.
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.
The appendix on the last sheet can be removed for reference for questions 2-4.

Questions

Question 1. Search [20 marks]

(a) [5 marks] Assume that the numbers in the figure below are the costs between two cities (nodes) in Romania. When *uniform cost search* is used and the problem is to get from Sibiu to Bucharest, provide the search path and the final solution.



Search Path: Sibiu → (Rimnicu Vilcea 80, Fagaras 99) → Pitesti (80+97 = 177) → Bucharest (99+211 = 310) → Bucharest (80+97+101 = 278)
Final solution: Sibiu → Rimnicu Vilcea → Pitest → Bucharest

(b) [3 marks] Briefly describe when *iterative deepening (depth-first) search* should be used in general in terms of the search space and the depth of solution.

In general, iterative deepening is the preferred uninformed search method when the search space is large and the depth of the solution is unknown.

(c) [7 marks] *Hill climbing* and *simulated annealing* are two search techniques.

(i) Describe the main idea of *hill climbing*. Draw a figure if necessary.

(ii) State a major limitation of *hill climbing*.

(iii) Briefly describe how *simulated annealing* is different from *hill climbing*.

(i) HC is a local search technique and aims to find the best state according to an objective function. It only keeps one state and its evaluation/performance, and choose the best successor. (ii) Can easily stuck to local maximum/optima. (iii) SA borrows the idea from mental annealing in physics by heating them to a high temperature then gradually cooling them. It is similar to HC, but with a random component attempting to jump from the local optimum.

(d) [5 marks] *Gradient descent search* and *(genetic) beam search* are two heuristic search methods. Briefly describe the differences between them in terms of

(i) whether they are local or global search techniques,

(ii) whether they are producing partial solutions or global solutions at each intermediate step, and

(iii) whether one or more solutions can be generated from each experiment run.

(i) GD is local; BS is global
(ii)GD: partial; BS: candidate
(iii) GD: one; BS: multiple

Question 2. Machine Learning Basics [30 marks]

(a) [4 marks] There are several different paradigms in machine learning. State the name of an algorithm or approach used in each of the following paradigms:

(i) Case based learning

(ii) Induction learning

(iii) Statistical learning

(iv) Connectionist learning

(b) [4 marks] In addition to a *training set* and a *test set*, a *validation set* is often used in (supervised) machine learning systems.

State the major role of the *validation set*.

(i) to avoid/control overfitting.
(ii) Both are used in the training process, but the training set is used to directly for training and extracting the pattern/classifier while the validation set is used for monitoring the training process to avoid overtraining/overfitting.

(c) [4 marks] Briefly describe the *K Nearest Neighbour method* used for classification tasks.

Each unseen instance (in the test set) is compared with all the instances in the training set to calculate the distance (typically Euclidean distance) or similarity for all the training instances,
Find the "nearest neighbour" (instance) from the training set based on some distance/similarity measures,
Then choose the class label of the nearest neighbour as the class label of the unseen instance in the test set.

(d) [6 marks] Suppose you are building a Naïve Bayes spam filter to distinguish *spam* messages from real email messages (*non-spam*). You have picked two key words: "discount", and "project" to characterise each message, and have counted how many of the messages contain each word:

	spam		**email (non-spam)**	
	word present	word not present	word present	word not present
"discount"	300	100	10	90
"project"	20	380	70	30
Total count	400		100	

If your spam filter was presented with a new message that contained both words "discount"and "project", would your spam filter classify the message as spam or as email (non-spam)? Show your working.
(Note: you do not need to use a "pseudo-count").

score(spam) = 300/400 * 20/400 * 400/500 = 3/100 = 15/500
score(non-spam) = 10/100 * 70/100 * 100/500 = 7/500
Therefore, it will choose spam.

(e) [8 marks] Consider the following data set describing 10 loan applications at a bank, of which 5 were approved and 5 were rejected. They are described by three attributes: whether the applicants have a job or not, whether their deposits are low or high, and whether their credit records are very good, good or bad.

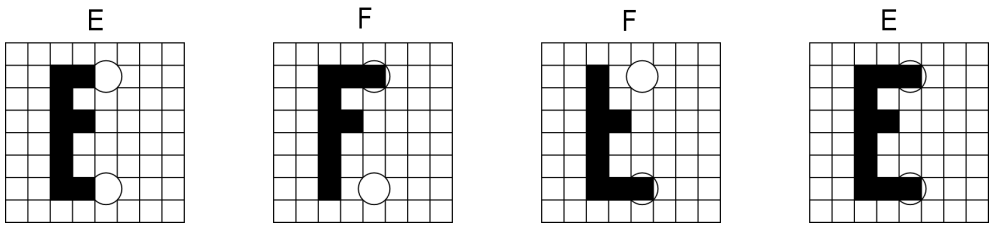
Instance	Job	Deposit	Credit	Class
1	true	low	very good	Approve
2	true	low	good	Approve
3	true	high	very good	Approve
4	true	high	good	Approve
5	false	high	very good	Approve
6	false	low	good	Reject
7	false	low	bad	Reject
8	true	low	bad	Reject
9	false	low	very good	Reject
20	false	high	bad	Reject

The bank wants to build a decision tree to help making loan-granting decisions. Which attribute should the bank choose for the *root* of the decision tree if they use the impurity function $p(\text{Approve})p(\text{Reject})$? Show your working.

Job: 5/10 * (4/5 * 1/5) + 5/10 * (1/5 * 4/5) = 4/25 = 16%
Deposit: 6/10 * (2/6 * 4/6) + 4/10 * (3/4 * 1/4) = 5/24 = 21%
Credit: 4/10 * (3/4 * 1/4) + 3/10 * (2/3 * 1/3) + 3/10 * (0/3 * 3/3) = 17/120= 14%)

Credit has the lowest score, therefore the algorithm will use Credit at the root

Peter Jackson used a perceptron (linear threshold unit) to solve a binary classification problem of four image instances as an “E” or an “F”:



In the perceptron, he used two input nodes corresponding to the two pixels marked with a circle in the figure and one output node corresponding to the output class label. The pixel values are either 0 (black) or 1 (white), and the desired output value is 0 for “E” and 1 for “F”. However, his perceptron could not converge no matter how he changed the learning parameters.

(f) [2 marks] Explain why Peter’s perceptron could not be trained successfully.

The instances are not linearly separable (by a hyperplane)
the perceptron learning algorithm can only classify instances that are linear separable

(g) [2 marks] Suggest two changes Peter could make that would enable the image instances to be learned successfully.

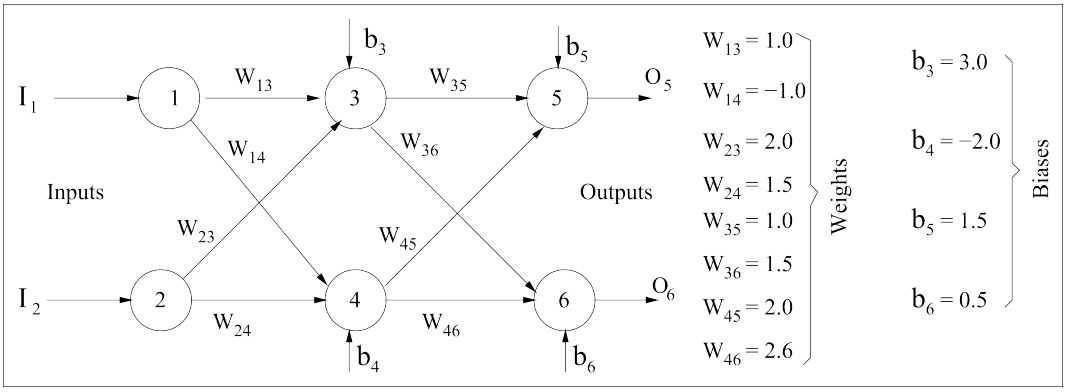
Possible ways: get better input features if possible; use a multilayer perceptron/neural network; use a better transfer function to replace the threshold function, use a better training algorithm such as the back propagation algorithm

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, and Support Vector Machines [20 marks]

(a) [9 marks] Consider the following feed forward neural network which 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 the vector (0.0, 0.0, 0.0, 0.0) for input and output nodes (the first two for inputs and the last two for outputs) and that the learning rate η is 0.25.

$O_1 = I_1 = 0; I_3 = b_3 = 3.0; O_3 = f(3.0) = 0.95; O_4 = f(-2.0) = 0.12$
$I_5 = 0.95 \cdot 1.0 + 0.12 \cdot 2.0 + 1.5 = 2.69; O_5 = f(2.69) = 0.93$
$\Delta W_{35} = \eta O_3 O_5 (1 - O_5) \beta_5 = 0.25 \times 0.95 \times 0.93(1 - 0.93)(0 - 0.93) = -0.0144$
$(W_{35})_{new} = (W_{35})_{old} + \Delta W_{35} = 1 - 0.0144 = 0.9856$

(b) [7 marks] Peter Smith has developed a classifier for distinguishing cancer cells from normal cells. He extracted 4 features from images of cells, used the standard multilayer feed-forward neural network, and applied the back propagation algorithm to train his network for classification. There are 500 examples in total from which he used 100 for network training and 400 for testing. The network architecture he used is 5-25-1. After training for 10,000 epochs, the network classifier obtained 99.5% accuracy on the training set, but only achieved 67% accuracy on the test set. Suggest three good ways to Peter for improving the (test) performance.

(1) re-split the data sets and use more examples for training; (2) use fewer hidden nodes; (3) Train fewer epochs; (4) any other reasonable suggestions such as get more and better features, or use a validation set to control overfitting.

(c) [4 marks] For a linear classifier, what does the term "margin" refer to, and how is it used in Support Vector Machines (SVMs)?

The margin is the distance from an input pattern to the decision surface, which in a linear classifier is a hyperplane. In SVMs we find the hyperplane that *maximizes the minimum margin*. [An additional point is that SVMs use the 'kernel trick' to do this in an implicit high dimensional feature space, without actually working in that space.]

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 [25 marks]

(a) [4 marks] State the representation for solutions, and search techniques, in Genetic algorithms and Neural networks:

(i) Genetic algorithms

- Representation:
- Search:

(ii) Neural networks

- Representation:
- Search:

(b) [6 marks] Briefly describe the general evolutionary process in Evolutionary Algorithms.

this should include initialisation, evaluation, selection, mating and when to stop.

(c) [4 marks] In evolutionary computation, *tournament selection* is a popular selection method. Briefly describe this method.

Tournament Selection: (1) for a given tournament size of n, this method randomly choose n individuals from the population and place them into the tournament. (2) With the tournament, the individuals compete against each other, and the best one based on the fitness is selected and placed into the mating pool for evolution/mating.

(d) [5 marks] Genetic Programming (GP) can be used for symbolic regression tasks.
In Assignment #2, you used GP to evolve 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 from the following data set (10 points).

x	-0.75	-0.50	-0.25	0.00	0.25	0.50	0.75	1.00	1.25	1.50
y	2.7227	1.5625	1.0977	1.0000	1.0352	1.0625	1.0352	1.0001	1.0977	1.5625

(i) Suggest a good terminal set.

(ii) Suggest a good function set.

(iii) Suggest a good fitness function.

(iv) Statistical regression searches a space of parameter values. Indicate the space that GP searches.

Terminal set: { X, R}, R is a random number
Function set: {+, –, *, %} or {+, -, ^ } or other reasonable sets
Fitness Function: mean squared error, sum squared error, absolute error, etc.
GP characteristics: (1) symbolic regression: GP automatically evolves the mathematical model and corresponding parameter/coefficient values. (2) GP does not need to assume any distribution of the data. (3) GP can evolve multiple models for a particular task using a single experiment run. (4) a small number of examples.
...

(e) [6 marks] The standard tree-based genetic programming approach has been applied to many classification tasks. In this approach, each evolved program typically returns a single floating point number. One of the key issues here is to use a *strategy* to translate the single output value of an evolved classifier program into a set of class labels.

(i) In Assignment #2, GP was used to evolve a classifier to categorise the 699 instances in the Wisconsin medical data set into either the *benign* class or the *malignant* class. Suggest a strategy (rule) for translating the single program output into the above two classes.

(ii) For multiple class classification problems, one simple method for this translation is the *program classification map*, which splits the program output space into predefined regions, each corresponding to a particular class. State two problems with this translation method, and suggest one method to overcome (or at least reduce) the problems.

(i) For binary classification, the natural translation would be: if the program output value is positive, then the instance associated with the inputs terminals is classified as class 1; otherwise, class 2.

(ii) Limitations/problems: the class boundaries are fixed; the boundaries need to be predefined; class orders are fixed.
Improvement methods: (1) decompose the multiclass classification problem into multiple binary classification problem, then use GP each for each binary classification subproblem; (2) use dynamic class boundary determination methods.

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. Philosophy of AI [10 marks]

Consider a process by which, one by one, the neurons in your brain are replaced by functional equivalents that are made from artificial materials but in all other respects behave exactly as biological neurons do - even to the level of adapting their connnections and thus taking part in learning.

As the process continues, more and more of your cerebral cortex gets replaced in this way. The interesting question arises: "At what point in this process do you stop being you?"

(a) [2 marks] Using one or two sentences, what is your position on this question? Characterize your position as dualist, monist / physicalist, or functionalist (or, if none of these fits, give another option).

Do the two answers match up?

(b) [8 marks] Give one argument in favour of your position, and one argument (such as another thought experiment) against it.

(a wide variety of answers are possible here - marks awarded for a coherent argument, and ability to see multiple sides of the issue, indicative of having given it some thought...)

Question 6. Reasoning under Uncertainty [10 marks]

Note: for questions that involve calculations, *show your working* to ensure maximum credit.

(a) [3 marks] One way of expressing the knowledge that two events X and Y are independent is to write

$$P(X|Y) = P(X)$$

Show how this implies that the following is also true:

$$P(X,Y) = P(X)P(Y)$$

...)

(b) [2 marks] Consider two Boolean variables A and B. If we know $P(B|A) = \frac{1}{3}$, which of the following do we also know?

- $P(B|\neg A)$ no
- $P(\neg B|A)$ yes
- $P(\neg B|\neg A)$ no

Bayes Rule provides a way to calculate how some data $\mathcal{D}$ should affect the degree of belief we should assign to some hypothesis $\mathcal{H}$, as follows:

$$P(\mathcal{H}|\mathcal{D}) = \frac{P(\mathcal{D}|\mathcal{H})\ P(\mathcal{H})}{P(\mathcal{D})}$$

(c) [3 marks] What are the usual names given to $P(\mathcal{H}|\mathcal{D})$, $P(\mathcal{D}|\mathcal{H})$ and $P(\mathcal{H})$?

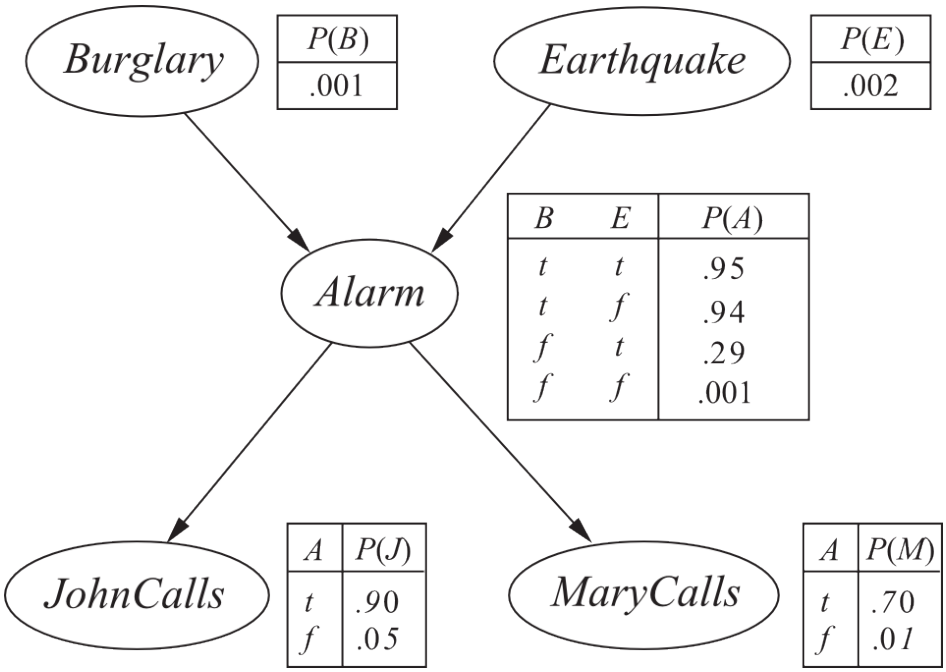
...

(d) [2 marks] How would you go about calculating the value of the denominator, $P(\mathcal{D})$?

...

Question 7. Belief Networks [20 marks]

Consider the following Belief Network, which represents two causes and two effects related to activation of a burglar alarm.



Each variable takes the value true (`t`) or false (`f`). We will abbreviate the variable names using their leading letters: B, E, A, J, and M.

(a) [3 marks] What is the probability $P(B = \texttt{t}, E = \texttt{t}, A = \texttt{t})$? That is, what is the probability that there is a *Burglary* **and** an *Earthquake* **and** the *Alarm* is triggered?

...

(b) [2 marks] What is the probability that $A = \texttt{t}$? (ie. what is the probability that the *Alarm* is triggered?)

...

(c) [3 marks] What is the probability that **both** *JohnCalls* **and** *MaryCalls*, if we know that the *Alarm* has been triggered?

...

(d) [2 marks] Draw the structure of the Belief Net corresponding to the following factorization:

$$P(A,B,C,D) = P(A)\ P(B)\ P(C|A,B)\ P(D|C)$$

...

(e) [5 marks] Recursively applying the "product rule" yields the general result that

$$P(x_1,\ldots,x_n) = \prod_{i=1}^{n} P(x_i|x_{1\ldots i-1})$$

(ignoring the slight intricacies of the $i = 1$ case), which places no additional constraints on the distribution over $x_1,\ldots,x_n$. By contrast, the factorization of a Belief Network is

$$P(x_1,\ldots,x_n) = \prod_{i=1}^{n} P(x_i|\text{Parents}_i)$$

(where Parents_i is the set of parents of node i), which *does* constrain the distribution. What does this difference imply about conditional independencies in Belief networks?

...

In the course we discussed the SUM-PRODUCT algorithm, which is also known as "belief propagation" because it takes the form of *message passing* on a *factor graph*.

(f) [2 marks] What quantity is the SUM-PRODUCT algorithm used to calculate?

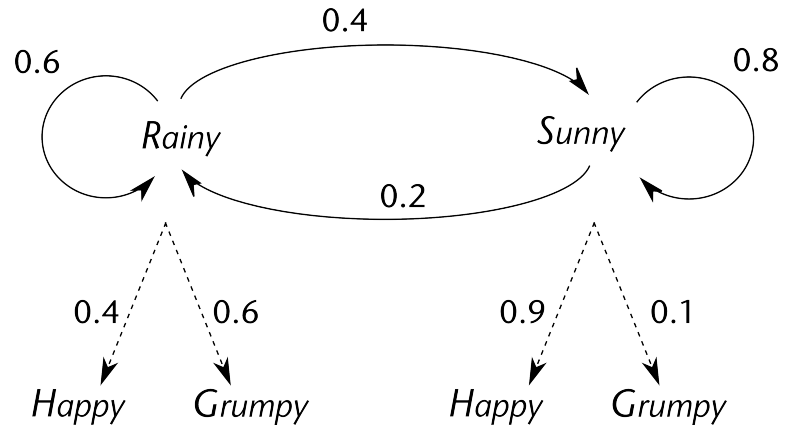
...

(g) [3 marks] In the SUM-PRODUCT algorithm, what are the basic operations carried out at the *factor* nodes?

...

Question 8. Reasoning about Sequences [18 marks]

Consider the following Hidden Markov Model, which represents the *Weather* over successive days (`Sunny` versus `Rainy`), and Sebastian's *Mood* on those same days (`Happy` versus `Grumpy`). The numbers represent conditional probabilities for transitions (solid lines) and for moods (dashed lines).



(a) [3 marks] Draw the corresponding factor graph structure that would result from "unrolling" this model over three successive days. Indicate which factors are "shared" in this network.

...

(b) [4 marks] If the probability of `Rainy` is 0.5 on day t (that is, $P(\mathit{Weather}_t = \texttt{Rainy}) = 0.5$), what is $P(\mathit{Weather}_{t+1} = \texttt{Rainy})$, assuming no other information is available? Show your working.

...

(c) [3 marks] Under the stationary distribution, what is the probability that *Weather* is `Rainy`, given these transition probabilities?

...

(d) [3 marks] Suppose that instead of knowing transition probabilities $(0.6, 0.4, 0.8, 0.2)$ you start off with no idea about the correct transition probabilities, but instead have observed a data sequence of *Weather* over 7 days:

```
Rainy, Sunny, Sunny, Rainy, Sunny, Sunny, Sunny
```

What values would you estimate for the transition probabilites, based on this data? (For full credit, you should employ Laplace Smoothing in arriving at an answer).

...

(e) [5 marks] *(Hard)* Suppose you have the same HMM structure and you don't know the transition probabilities. Outline how you could arrive at sensible transition probabilities even if you *never* observed the weather, but did observe a sequence of "moods".

...

Question 9. Planning [27 marks]

In Classical Planning, two ways of deriving a plan are known as `forward` and `backward` chaining.

(a) [4 marks] Outline how the two components of Action Schemas are used to accomplish `forward` chaining.

...

(b) [5 marks] Describe a scenario in which we would expect `backward` chaining to be much more efficient than `forward` chaining, and explain why this is so.

...

(c) [4 marks] A Markov Decision Process (MDP) is defined by which four quantities?

- states, $s_1, s_2, \ldots$
- actions, $a_1, a_2 \ldots$
- rewards for each state, $r_1, r_2 \ldots$
- transition probabilities $P_{s'|s,a}$

(d) [3 marks] What is meant by a "policy" for an MDP?

...

(e) [3 marks] What is the long-term value $V(s)$ of a state s in an MDP, in terms of the future rewards R_t, where a discounting factor γ is used?

...

The "Back-Up" equation is a recursive equation describing the value of a state s in terms of the values of states s' one step in the future:

$$V^{\pi}(s) = R(s) + \gamma \sum_{a} \pi_{a|s} \sum_{s'} P_{s'|s,a} V^{\pi}(s')$$

(f) [4 marks] In words, explain why there is a sum over a and over s' in this equation.

...

(g) [4 marks] Value Iteration *assumes* an optimal policy in order to find the optimal value function, $V^{\star}$. What effect does the optimal policy have on the above equation?

...

****

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