



**EXAMINATIONS – 2015
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 removed for reference for questions 2-4.

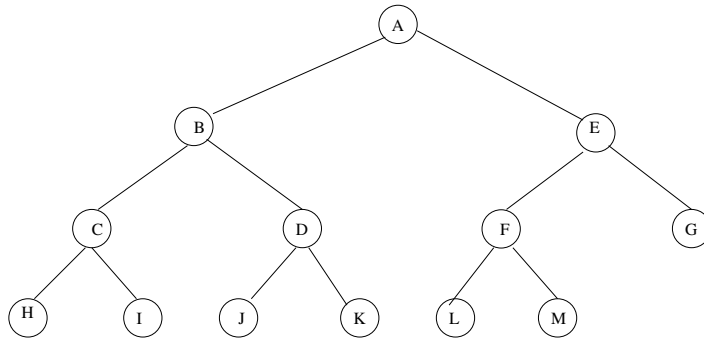
Questions

- 1. Search [15]
- 2. Machine Learning Basics [20]
- 3. Neural Networks [10]
- 4. Evolutionary Computation and Learning [15]
- 5. Representing Uncertainty [20]
- 6. Naive Bayes, Generative Models, and EM [10]
- 7. Modelling Sequences [10]
- 8. Making Decisions, Planning, etc. [20]

Question 1. Search

[15 marks]

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



(a) [2 marks] Assuming that you are using *depth first* search, state the search order/path using the letters in the nodes.

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

(c) [4 marks] *Hill climbing* is a basic local search technique.

- (i) Describe the main idea of this technique. Draw a figure if necessary.
- (ii) State a major limitation of this technique.
- (iii) State a solution to avoid (or at least reduce the degree of) the limitation in part (ii).
- (iv) Which of the search spaces, *state space* or *solution space*, is this technique searching for?

(d) [5 marks] *Gradient descent search* and (*genetic*) *beam search* are two heuristic search methods.

- (i) State a machine learning paradigm/technique that uses each of the two methods.
- (ii) In his experiments, Peter Smith found that *gradient descent search* performed better than *genetic beam search* for his problem. So he claimed that gradient descent search is a *global* search technique and genetic beam search is a *local* search technique. Do you agree with Peter's claim? Justify your answer.

(d) [6 marks] Consider the following data set describing 10 menu items from a restaurant, of which 5 are popular with customers, and 5 are not. They are described by three attributes: whether they are spicy or mild, whether they have sauce or not, and whether the protein base is vegetarian, beef, or chicken.

| Spice | Sauce | Protein | Category |
|-------|-------|------------|-----------|
| spicy | yes | beef | Popular |
| spicy | yes | beef | Popular |
| spicy | yes | vegetarian | Popular |
| mild | no | beef | Popular |
| spicy | no | beef | Popular |
| spicy | yes | chicken | unpopular |
| spicy | yes | vegetarian | unpopular |
| spicy | no | beef | unpopular |
| mild | no | beef | unpopular |
| mild | no | chicken | unpopular |

Which attribute would the decision tree building algorithm choose for the root of a decision tree for predicting whether a menu item will be popular or unpopular? Show your working.

(e) [4 marks] John Smith used a perceptron (linear threshold unit) to solve a binary classification task with the following labelled instances:

| Input Feature 1 | Input Feature 2 | Input Feature 3 | Output Class |
|----------------------------|----------------------------|----------------------------|-------------------------|
| 0 | 1 | 0 | 0 |
| 0 | 1 | 1 | 1 |
| 1 | 1 | 0 | 1 |
| 1 | 1 | 1 | 0 |

His perceptron used three input nodes and one output node, corresponding the three input features and the output class. It was trained using the usual perceptron learning rule (that we discussed during the lectures), but the weights did not converge no matter how he changed the learning parameters.

- (i) Explain why John's perceptron was not successful.
- (ii) Suggest two improvements that would enable the instances to be learned successfully.

Student ID:

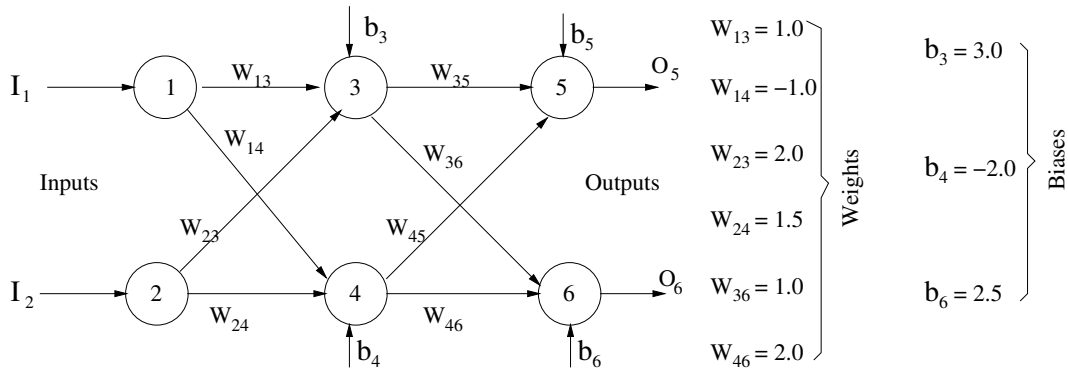
SPARE PAGE FOR EXTRA ANSWERS

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) [7 marks] Consider the following feed forward neural network which uses the sigmoid/logistic transfer function (see Appendix B),



- (i) What will be the output of node 6 (O_6) for the input vector (0.0, 0.0)?
- (ii) What will be the new value of weight w_{46} 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) corresponding to the two input feature values and the two target output values, and that the learning rate η is 0.2.

Show your working.

(b) [3 marks] John Smith has developed a classifier for distinguishing *cancer* cells from *normal* cells. He extracted 5 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 1000 examples in total from which he used 100 for network training and 900 for testing. The network architecture he used is 5-30-2. After training for 20,000 epochs, the network classifier obtained 99.9% accuracy on the training set, but only achieved 56% accuracy on the test set. Suggest three good ways to John 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) [4 marks] The *genetic algorithm* is one of the techniques in evolutionary computation and learning. State six additional techniques in evolutionary computation and learning.

(b) [5 marks] In the context of genetic algorithms and genetic programming, briefly explain why the *mutation* operator is usually needed (in addition to crossover) and why it is only set to a small rate (compared to the crossover operator).

(c) [6 marks] Genetic Programming (GP) is considered a good method for symbolic regression tasks. In assignment 2, you used GP to evolve a mathematical function to reveal the relationship between the output variable and the input variable(s) from a (training) set of instances.

Assuming your task is to use GP to map a single input variable x to the single output variable y from the following data set (20 points):

| | | | | | | | | | | |
|---|--------|--------|--------|-------|-------|-------|-------|-------|--------|--------|
| x | -2.0 | -1.75 | -1.50 | -1.25 | -1.00 | -0.75 | -0.50 | -0.25 | 0.00 | 0.25 |
| y | 37.000 | 24.160 | 15.062 | 8.910 | 5.000 | 2.722 | 1.562 | 1.097 | 1.000 | 1.035 |
| x | 0.50 | 0.75 | 1.00 | 1.25 | 1.50 | 1.75 | 2.00 | 2.25 | 2.50 | 2.75 |
| y | 1.065 | 1.035 | 1.000 | 1.097 | 1.562 | 2.727 | 5.000 | 8.912 | 15.062 | 24.160 |

- (i) Suggest a good terminal set.
- (ii) Suggest a good function set.
- (iii) Suggest a good fitness function.
- (iv) Briefly describe the main differences between the *statistical regression* method and the *GP for symbolic regression* method.

Student ID:

SPARE PAGE FOR EXTRA ANSWERS

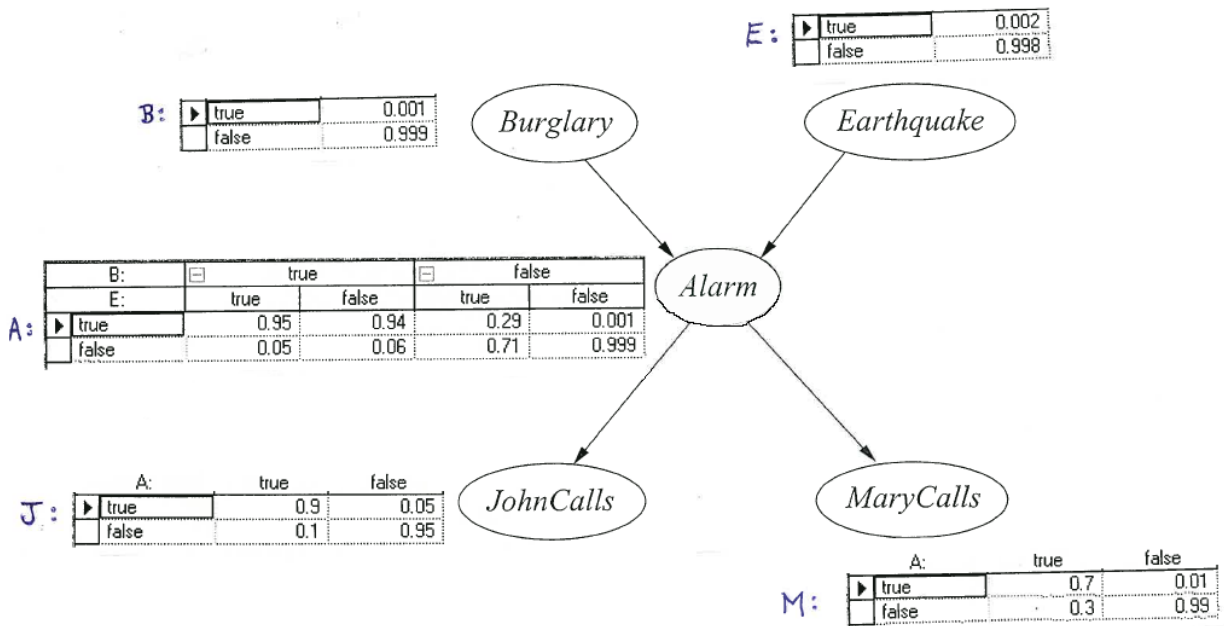
Cross out rough working that you do not want marked.
Specify the question number for work that you do want marked.

Question 5. Representing Uncertainty

[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` or `false`. You can abbreviate the variable names using their leading letters: *B*, *E*, *A*, *J*, and *M*.

(Note regarding calculations: it doesn't really matter whether you compute the exact final numbers here - the main thing is to demonstrate that *you know how* to calculate them, so **show all working**).

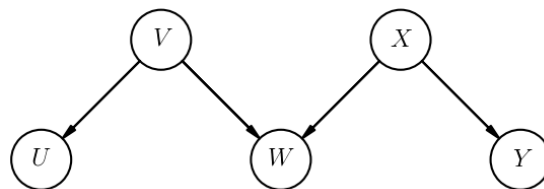


(a) [3 marks] What is the probability $P(B = \text{true}, E = \text{false}, A = \text{false})$? That is, what is the probability that there is a *Burglary* but no *Earthquake*, and the *Alarm* fails to be triggered?

(b) [2 marks] Show that J and M become independent once we know A , *i.e.*, prove that

$$P(J, M|A) = P(J|A) P(M|A)$$

(Note: this is true regardless of the particular values of the CPTs).

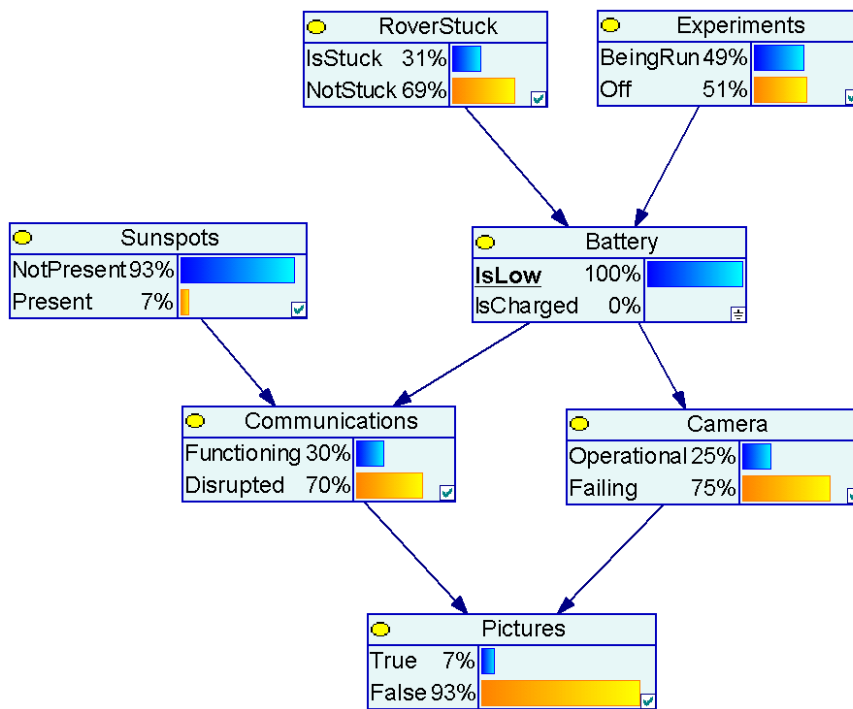


(c) [2 marks] Write an equation giving the factorisation of the joint probability $P(U, V, W, X, Y)$ implied by the network shown above.

(d) [2 marks] For the network shown above, are the following statements true, or false?
Reminder: the \perp symbol stands for “independent”, and the vertical line stands for “given”.

- $V \perp Y \mid X$

- $U \perp Y \mid W$



(e) [2 marks] In the belief network shown above, the state of the `Battery` variable has been observed to be `IsLow`, but none of the other variables have been observed. Which nodes would have their belief distributions changed if, in addition, you were now to discover the state of `Sunspots`?

(f) [2 marks] For the network shown above, if all nodes take only two values (e.g. Booleans), how many free parameters are required in order to fully specify the joint distribution?

(g) [4 marks] What is a *factor graph*, and how does it relate to the structure of a belief network? It may help to draw an example of each that shows the similarities and differences in their structure.

The “sum-product algorithm”, otherwise known as “belief propagation”, is an algorithm that runs on the factor graph. The algorithm takes the form of message passing.

(h) [1 mark] What is the message generated by a variable node that is a leaf (*i.e.* has no child nodes) and has *not* been observed?

(i) [2 marks] In the course of the algorithm, when may a node send a message along a given link?

Question 6. Naive Bayes, Generative Models, and EM

[10 marks]

In a *Naive Bayes classifier*, a vastly simplified network is used in place of the full joint distribution over all the variables.

(a) [2 marks] A spam filter that is a Naive Bayes classifier is essentially making very strong assumptions about the data. What are those assumptions?

(b) [2 marks] The Naive Bayes classifier often works well even in cases where its assumptions are not strictly obeyed. Why do you think this might be?

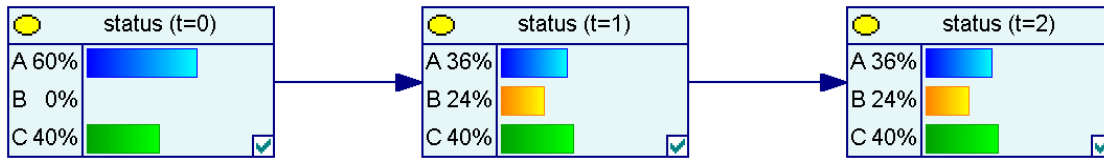
(c) [4 marks] What is the key difference between the discriminative and the generative approaches to classification? Give an example of each (ie. an algorithm, architecture, or similar).

A belief network can be “trained” from a data set by the EM algorithm, even when numerous values in the data are missing / unknown.

(d) [2 marks] Why do the E and M steps need to be iterated several times instead of once each?

Question 7. Modelling Sequences

[10 marks]



(a) [4 marks] The belief net shown above is a Markov network, which has been “unrolled” for 2 transitions. No nodes have been observed, and the numbers indicate the outcome of running belief propagation. Write down a CPT (conditional probability table) for the transition factor, $P(\text{status}_{t+1} | \text{status}_t)$, which would lead to the belief values shown.

(b) [3 marks] What is the *most likely sequence* of states:

$$\text{status}(t = 0) \rightarrow \text{status}(t = 1) \rightarrow \text{status}(t = 2)?$$

(c) [3 marks] When would a Particle Filter approach be better than a Hidden Markov Model? Give an example.

Question 8. Making Decisions, Planning, etc.

[20 marks]

A *Markov Decision Process* (MDP) is defined by the following four quantities:

- states, s_1, s_2, \dots
- actions, a_1, a_2, \dots
- rewards for each state, R_1, R_2, \dots
- transition probabilities $P_{s'|s,a}$

(a) [2 marks] What is a “policy” for an MDP?

(b) [4 marks] What is a “policy” for a POMDP? (*Note:* POMDP stands for Partially Observable MDP). In particular, how does it differ from the case of an MDP, and what are the implications?

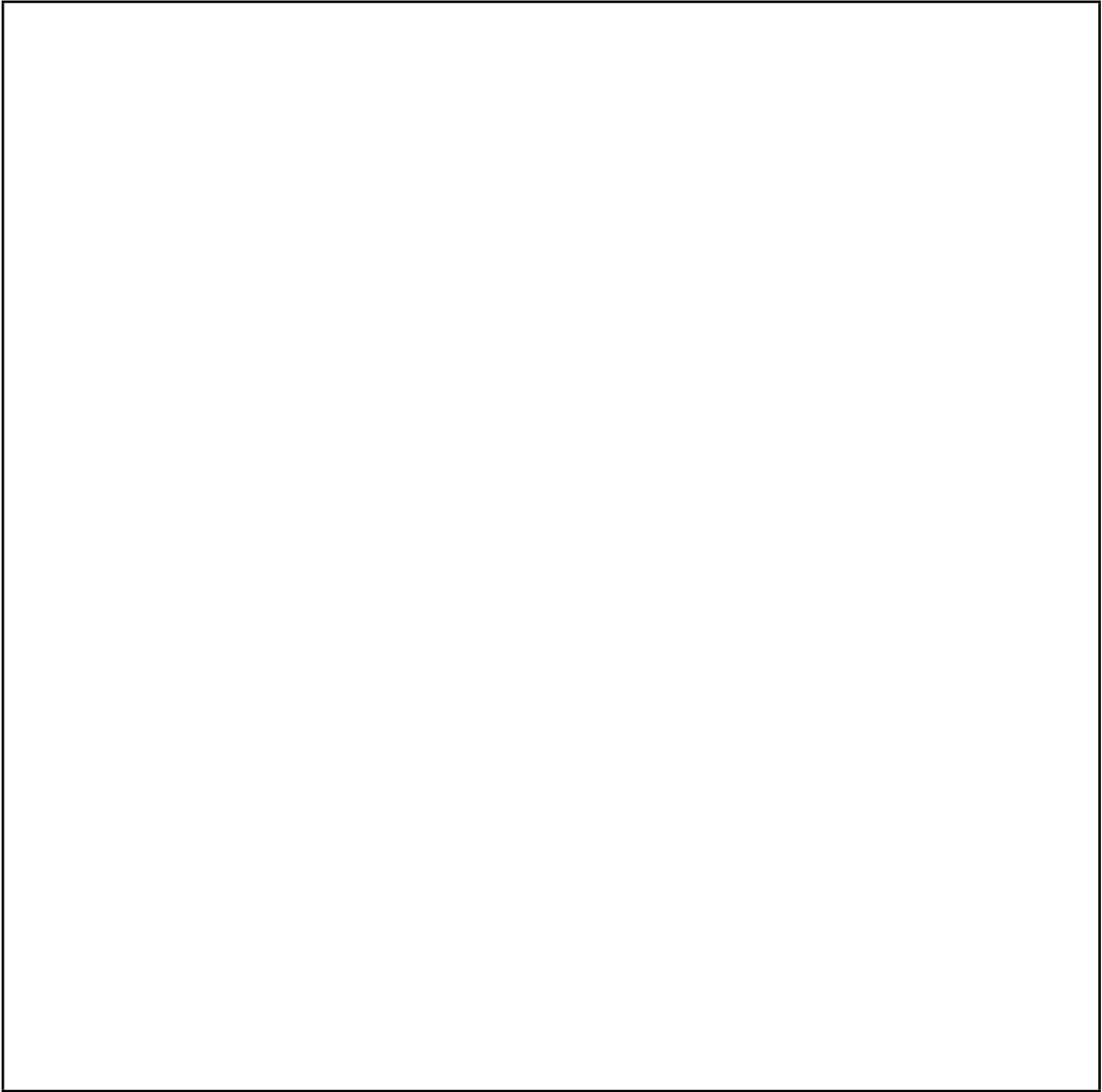
In an MDP, the “value” U_s of a state s is defined to be the expected sum of future rewards R , where each successive reward is *discounted* according to how far it is into the future. The algorithm *Value Iteration* can find the *optimal* value function (for each state s), which we might denote U_s^* .

(c) [5 marks] How should a *rational* agent make use of the optimal value function?

Classical planning relies on deterministic outcomes to remain tractable for large problems. In Classical Planning, two ways of deriving a plan are known as *forward* and *backward* chaining.

(d) [5 marks] Give a scenario in which *forward* chaining would be preferable to *backward* chaining, AND a scenario in which the reverse is true.

(e) [4 marks] Outline the distinctions between the following three perspectives on the mind-body problem: Dualism, Physicalism, and Functionalism. *Note: this is only worth 4 marks, so answer at an appropriate depth.*



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.

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

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

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

$$O_i = f(I_i) = f\left(\sum_k w_{k \rightarrow i} \cdot o_k + b_i\right) \quad (3)$$

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

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

$$\beta_j = d_j - o_j \quad (6)$$

B Sigmoid/Logistic Function

