UNIVERSITY OF OTAGO EXAMINATIONS 2015

COMPUTER SCIENCE

COSC343

ARTIFICIAL INTELLIGENCE

(TIME ALLOWED: THREE HOURS)

This examination comprises 7 pages.

Candidates should answer questions as follows:

Candidates should answer **5** questions.

Questions are worth 12 marks and sub marks are shown thus: (2)

The total number of marks for this exam is 60.

The following material is provided:

Nil.

Use of calculators

No restriction on the model of calculator that may be used, but no device with communication capability shall be accepted as a calculator. Calculators are subject to inspection by the examiners.

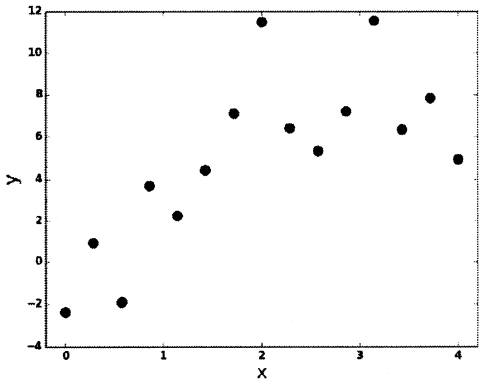
Candidates are permitted copies of:

Nil.

TURN OVER

1. Machine learning

The following graph presents a small set of points sampled from an unknown polynomial function $y = f(x)$. Measurements of y include a component of Gaussian noise.



(a) To estimate the function f, a supervised learning algorithm needs a **hypothesis space** to explore, as well as a set of training data points. What is a hypothesis space? (1)

(b) Suppose we adopt the set of **polynomial functions of order 2** as our hypothesis space.

(i) What are the two dimensions of this hypothesis space? (1)

(ii) What does each point in the hypothesis space represent? (1)

(iii) Explain how the training data points can be used to define an **error surface** for this hypothesis space. Include the concept of **sum squared error** in your answer. (3)

(c) Using the notion of an error surface, explain informally how the **linear regression** technique estimates a polyomial function from a set of training data points. (2)

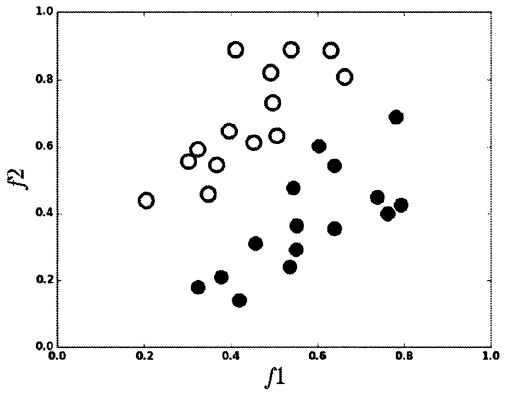
(d) Suppose we perform two regression analyses: one using polynomials of order 2 as its hypothesis space, and another using polynomials of order 3.

(i) Which analysis is likely to minimise error on the above training set? Explain your answer. (2)

(ii) As the hypothesis space is increased, the danger of **overfitting** the training data grows. How can we check to make sure that a supervised learning algorithm isn't overfitting its training data? (2)

2. Perceptrons

The following graph presents some training data for a **classification** algorithm. Filled circles represent instances of Class A; empty circles represent instances of Class B. For each training instance, the values of two numerical features $f1$ and $f2$ are recorded. Each feature is scaled within the interval [0, 1].



(a) You are asked to design a single-unit **perceptron** which can be trained to classify items as Class A or Class B, based on their $f1$ and $f2$ values, using the above data as training data.

(i) To train the perceptron, you will need to **encode** the classes A and B as numbers in the interval [0, 1]. Choose an encoding for classes A and B, and write it down. (1)

(ii) Write down a sensible **activation function** for your perceptron unit. (1)

(iii) Sketch a diagram showing the structure of your perceptron, including inputs and outputs. Don't forget to include a bias unit. (2)

(iv) The first instance of Class B in the training data has $f1$=0.2 and $f2$=0.41. Describe how your perceptron's **error** on this training instance is computed, as a function of its weights. (2)

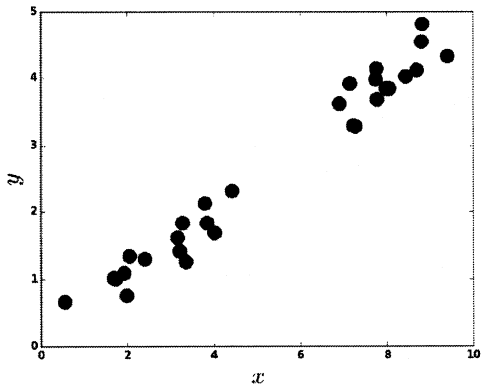
(v) Describe how the **perceptron learning rule** will *adjust* these weights based on the error for this training instance. (2)

(b) The training data shown above has an attractive property, which allows a single-unit perceptron to identify the class of each instance perfectly, without making any errors. Describe this property, explaining any terms you use. (2)

(c) Sketch a graph showing another set of training data for classes A and B, for which a single-unit perceptron would *not* be able to learn a perfect classification. Explain why the single-unit perceptron would fail on your training set. (2)

3. Unsupervised learning

The following graph shows a set of unlabelled data in two dimensions x and y.



(a) Explain informally what **principal component analysis** (**PCA**) is, and how it could be used to re-represent the above data. Include a diagram sketching the results of PCA, to show its effects. (3)

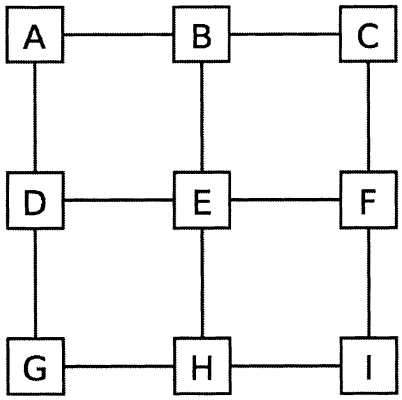
(b) Provide pseudocode for the **k-means clustering** algorithm, and explain how this algorithm could be used to find structure in the above data. (3)

(c) Explain how an **autoencoder** neural network with a single unit in its hidden layer could be used to re-represent the above data. In your answer you should refer to the structure of the network, and also describe the method by which it is trained. (4)

(d) How does the representation learned by the autoencoder compare to the representation produced by PCA? (2)

4. Search

The NSA deploys a spy robot in Manhattan, disguised as a businessman. The robot patrols the grid of streets shown below. It is currently at location B, and its current goal is to navigate to location I. The robot uses **A* search** to compute a path to its goal.



(a) In the A* algorithm, each node created during the search is associated with a **path cost** and a **heuristic value**. Define these two concepts, and explain how they are used in A* search. (3)

(b) To assign a heuristic value to a node, the robot uses the **Manhattan distance** between the node's associated state and the goal state. The Manhattan distance between two locations is the sum of the number of blocks to be traversed horizontally and vertically to get from one location to the other. (Thus the distance between locations A and D is 1, and the distance between A and H is 3.) Show that the Manhattan distance is an **admissible** heuristic. (2)

(c) Assume that the ith node created during the search is labelled N_i, and has fields (l_i, pc_i, h_i) representing its associated location, path cost and heuristic value respectively. The first node created, representing location B, is $N_1(B, pc_1, h_1)$.

(i) Using this notation, show the contents of the **fringe** of nodes after each iteration of the A* algorithm until the node representing the goal location is selected from the fringe. (If two nodes have the same value, they are ordered on the fringe alphabetically by location: for instance $N_i(A\ldots)$ precedes $N_j(B\ldots)$.) (4)

(ii) Draw the search tree produced by the search algorithm. For each node in the tree, specify its associated location. (3)

5. Probabilistic reasoning

After a mysterious epidemic, 5% of the population of New Zealand become **zombies**: apparently ordinary citizens who fly into a murderous rage on the full moon. In an effort to detect zombies, the University of Otago develops a test, called **Test A**, that returns a verdict of 'positive' or 'negative'. In trials on known zombies, Test A is found to be 80% accurate. Specifically:

$$P(test\_A\_positive|zombie) = 0.8$$
$$P(\neg test\_A\_positive|\neg zombie) = 0.8$$

Students at the University are screened using Test A. Jim, a computer science student, is upset to learn he has tested positive by Test A.

(a) Use Bayes' rule over distributions to calculate the probability that Jim is a zombie, given that he tested positive. (Don't forget to take into account the **prior probability** of zombies in the NZ population.) (5)

(b) The University runs a second test on Jim, **Test B**, that screens for different symptoms. This test is 90% accurate for its positive diagnoses, and 70% accurate for its negative diagnoses.

(i) Explain informally why the results of Tests A and B are not **fully independent** of one another. (2)

(ii) What assumption would we make about the independence of the two tests in a **naive Bayes** probability model? (1)

(iii) On the basis of this assumption, draw a **Bayesian network** for the variables $Zombie$, $Test\_A\_positive$ and $Test\_B\_positive$. (2)

(iv) Jim returns positive for Test B as well. Use your Bayesian network to calculate the probability that Jim is a zombie given both his test results. (2)

6. Natural Language

Consider the following sentence, from the novel *Pride and Prejudice*.

You danced with the handsomest girl in the room.

And the following grammar:

S	→	NP, VP	Pron	→	you
S	→	S, PP	Det	→	the
NP	→	Pron	N	→	girl
NP	→	Det, N	N	→	room
NP	→	Det, Adj, N	Adj	→	handsomest
NP	→	NP, PP	V0	→	danced
VP	→	V0	P	→	in
PP	→	P, NP	P	→	with

(a) The grammar can assign two different structures to the sentence. Draw the two structures, and specify in words the meaning of each, making clear the differences between them. (4)

(b) The next sentence in the novel begins as follows:

Do let me ask my partner to ...

(i) Explain how we could use the text of *Pride and Prejudice* to build a **probabilistic language model** which could make sensible predictions about the next word in this sentence. (3)

(ii) Explain how we could use text of *Pride and Prejudice* to train a **simple recurrent network** to make predictions about the next word. (3)

(iii) What problems will both these predictive models face? (2)

END