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

EXAMINATIONS — 2005

MID-YEAR

COMP 422

DATA MINING, NEURAL NETS AND GENETIC PROGRAMMING

Time Allowed: 3 Hours

Instructions: Attempt ALL Questions.

The exam will be marked out of 180.
Non-programmable calculators are permitted.
Graph paper is provided.
Foreign language dictionaries are permitted.
Some useful formulae and functions can be found in Appendices A and B on pages 8-9.

Questions

	Marks
1. Data Mining and Knowledge Discovery	[35]
2. Computer Vision and Image Processing	[20]
3. Performance Evaluation	[15]
4. Neural Networks	[45]
5. Genetic Programming	[50]
6. Learning Theory	[15]

Question 1. Data Mining and Knowledge Discovery [35 marks]

This question concerns the basic concepts and algorithms of data mining (DM) and knowledge discovery in databases (KDD).

(a) [5 marks] Describe the terms *knowledge discovery in databases*, and *data mining*.

(b) [10 marks] Describe the KDD process.

(c) [9 marks] DM has a number of common tasks. Briefly describe each of the following DM tasks and give one example for each of them:

(i) Classification

(ii) Regression

(iii) Link analysis

(d) [5 marks] Briefly describe the *nearest neighbour* method for classification.

(e) [6 marks] Neural networks and genetic programming are two commonly used methods in data mining. Briefly describe the main differences between them in terms of representations and search techniques.

Question 2. Computer Vision and Image Processing [20 marks]

(a) [3 marks] Briefly describe the main differences between *computer vision* and *image processing*.

(b) [3 marks] List one application for each of the three image operators: *and*, *subtraction* and *multiplication*.

(c) [9 marks] There are six main aspects of image analysis: preprocessing, edge detection, segmentation, transformation, feature extraction and pattern classification. Briefly describe each of them.

(d) [5 marks] Briefly describe the *template matching* approach for finding locations of the small objects in large images.

Question 3. Performance Evaluation [15 marks]

(a) [9 marks] Assume that a classifier is applied to an object classification problem. There are two classes in the problem: *class1* and *class2*. The test set has 700 *class1* objects and 300 *class2* objects. At a confidence threshold level of 0.55, the classifier reports 600 objects for *class1* (of which 560 are correct). All the given objects are classified as either *class1* or *class2* by the classifier.

(i) Calculate the overall accuracy of the classifier.

(ii) Calculate the TPF (true positive fraction), the FPF (false positive fraction), the TNF (true negative fraction), and the FNF (false negative fraction) for *class1*.

(iii) Draw an ROC curve for *class1* using the results obtained in part (ii) and the additional results from the following table:

Threshold	**0.50**	**0.60**	**0.70**	**0.80**
TPF(%)	70	85	90	95
FPF(%)	5	30	50	60

(b) [6 marks] A detection system is used for finding all the objects of interest in a large image. Suppose:

- the image size is 1020×1020 pixels,
- the number of objects of interest in the image is 100,
- the size of the biggest object is 16×16 pixels, and
- a detector with an input field of 20×20 pixels is used as a template, and applied in a moving window fashion, over the large image to locate the objects of interest. This is done by scanning the image from the top-left corner, across and down, pixel by pixel.

At a confidence threshold level of 0.91, 250 objects are reported by the detection system, of which 75 are correct.

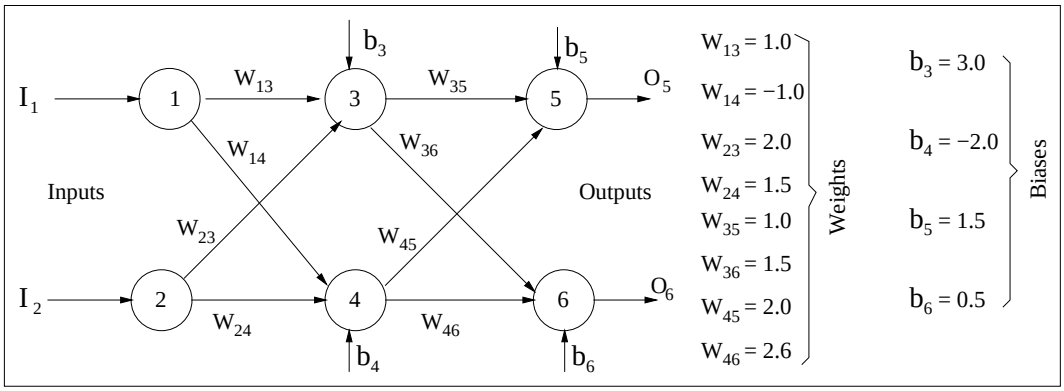
(i) Calculate the detection rate (DR) and false alarm rate (FAR) at this confidence threshold.

(ii) Calculate the precision and recall at this confidence threshold.

(iii) Assuming that only the object centres represent the correct locations of the objects, calculate the true positive fraction (TPF) and false positive fraction (FPF) at this confidence threshold.

Question 4. Neural Networks [45 marks]

(a) [12 marks] Consider the following feed forward network which uses the sigmoid/logistic transfer function (see Appendix B),



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

(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) and that the learning rate η is 0.3. Show your working.

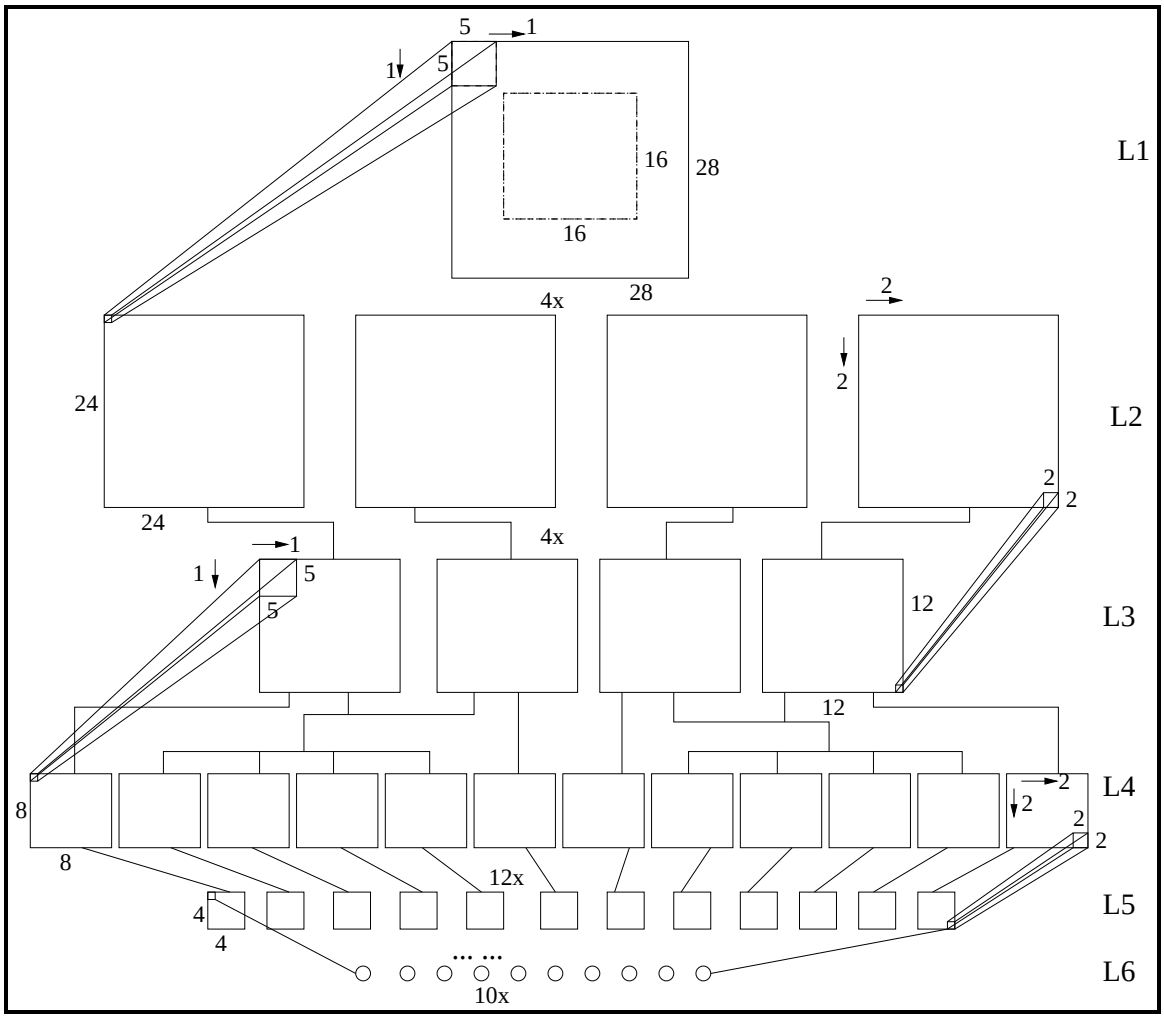
(b) [9 marks] John Smith has developed a neural network classifier to solve a prediction problem which involves the feature *time* of the 24 hour clock of a day. When he developed the classifier, he used the integer hours of a day (e.g. 1, 2, ..., 24) to represent the variable *time* and fed these values to a single input node of his neural network when he trained the network. However, the results obtained based on this representation were very bad, particularly for times near *midnight*.

State the main problems of this representation and suggest what changes John could make to improve the representation of `time` as input to the neural network. Draw a figure to show your suggestion if necessary.

(c) [8 marks] *Weight smoothing* and *centred weight initialisation* can be used to improve the performance of multilayer feed forward networks for object recognition. Briefly describe the main ideas of these two methods.

(d) [6 marks] Kohonen self-organising maps are often used for solving clustering problems. Briefly describe the architecture and the main idea of this method for clustering. Draw a figure if necessary.

(e) [10 marks] Consider the following shared weight network architecture:



with the following definitions:

```
% LeNet network

 ... ... ... ... (Omitted)

% L3
TYPE=SUBSAMPLING_MAP
SIZE=4@12x12 FILTER=2x2>2,2 CONNECTED=1000,0100,0010,0001

% L4
TYPE=FEATURE_MAP
SIZE=12@8x8 FILTER=5x5>1,1 CONNECTED=1000,1100,1100,1100,1100,0100,0010,0011,0011,0011,0011,0001

... ... ... ... (Omitted)
```

(i) Calculate the number of nodes in layer L4,

(ii) Calculate the total number of parameters (links+biases) between layers L3 and L4, and

(iii) Calculate the number of free parameters (weights+biases) between layers L3 and L4.

Question 5. Genetic Programming [50 marks]

(a) [8 marks] Assume that we use a genetic programming system to automatically evolve programs for regression problems such as fitting the following function or others of a similar form:

$$f(x) = \begin{cases} x^2 + \log_{10} x - x^4 & , \quad x \geq 0 \\ \cos x + \frac{1}{x} - 3.2 & , \quad x < 0 \end{cases}$$

(i) Determine an appropriate terminal set.

(ii) Determine an appropriate function set.

(iii) Suggest a good fitness function.

(iv) Determine the fitness cases.

(b) [6 marks] Describe the differences between genetic programming and genetic algorithms.

(c) [6 marks] Describe the differences between strongly typed genetic programming and standard genetic programming.

(d) [9 marks] To improve the standard crossover operator in genetic programming, a number of methods have been developed. These include the *brood recombination* method [Tackett 1994], the *intelligent crossover* method [Teller 1996, Iba 1996], and the *explicit defined introns* method [Nordin 1996]. Briefly describe the main idea behind each of these three methods.

(e) [15 marks] In standard tree based genetic programming for classification tasks, each evolved program typically returns a single floating point number, which is translated into a set of class labels. For classification problems with three or more classes, one simple method for this translation is the *program classification map* (or *static range selection*), which splits the program output space into predefined regions, each corresponding to a particular class. However, this method has a number of limitations. For example, the boundary values are fixed and need to be predefined, and the ordering of the classes is also fixed.

To overcome these limitations and improve the system performance, a number of improvements can be considered for multi-class classification problems. Remaining in the tree based genetic programming paradigm, suggest an improvement and discuss it in terms of:

(i) Class translation rules,

(ii) Program structure, and

(iii) Search techniques.

(f) [6 marks] Consider using genetic programming for finding objects of interest in large images. Suppose that the objects of interest belong to three different classes.

In the training process, each evolved program is used as a template, in a moving window fashion, to locate and classify the objects of interest in the large images. Each genetic program uses a squared input field as input which is large enough to contain every single object and from which the features are computed as terminals of the genetic programming system. A set of arithmetic operators and a conditional operator constitute the function set.

In each of the following questions, choose your answer from standardised fitness functions.

- **(i)** If the goal were to find all the objects of interest and not to care much about the number of false alarms, suggest a good fitness function.
- **(ii)** If the goal were to achieve high recall and high precision, suggest a good fitness function.
- **(iii)** If the goal were to achieve high recall and high precision, and to evolve short programs for solving this problem, suggest a good fitness function.

Question 6. Learning Theory [15 marks]

(a) [10 marks] Assume that c is the true classifier, h is the classifier we have learned using a learning algorithm L, and x is an individual example in the training set.

Given an accuracy measure ϵ and a confidence measure δ, describe the main ideas of the PAC learning model for evaluating whether the classifier (h) is "good" and whether the learning algorithm (L) is "good".

(b) [5 marks] In the PAC learning model, the VC (Vapnik-Chervonenkis) dimension is a measure of the learning capacity of a learning system and is often used to predict the number of training examples required. For a multilayer feed forward neural network with an architecture of 90-20-10 (90 input nodes, 20 hidden nodes and 10 output nodes):

- **(i)** Calculate the lower bound of the VC dimension for the neural network.
- **(ii)** If we want to achieve an accuracy level of 99% and a confidence level of 99.9%, what prediction does the PAC model with the VC dimension give for the number of examples needed to train the network?

A Some Formulae You Might Find Useful

$$P(h_j|x_i) = \frac{P(x_i|h_j)P(h_j)}{P(x_i)} \tag{1}$$

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

$$O_i = f(I_i) = f(\sum_j w_{j\to i} \cdot o_j + b_i) \tag{3}$$

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

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

$$\beta_z = d_z - o_z \tag{6}$$

$$\Delta w_{i\to j}(t) = \eta o_i o_j (1-o_j)\beta_j + \alpha \Delta w_{i\to j}(t-1) \tag{7}$$

$$TSS = \frac{1}{2}\sum_{p=1}^{n}\sum_{i=1}^{m}(t_{pi} - o_{pi})^2 \tag{8}$$

$$MSE = \frac{TSS}{n} = \frac{1}{2n}\sum_{p=1}^{n}\sum_{i=1}^{m}(t_{pi} - o_{pi})^2 \tag{9}$$

$$RMSE = \sqrt{\frac{2TSS}{n \cdot m}} \tag{10}$$

$$Recall = \frac{\sum_{j=1}^{n}\sum_{i=1}^{m} N_{true}(i,j)}{\sum_{j=1}^{n}\sum_{i=1}^{m} N_{known}(i,j)} \times 100\% \tag{11}$$

$$precision = \frac{\sum_{j=1}^{n}\sum_{i=1}^{m} N_{true}(i,j)}{\sum_{j=1}^{n}\sum_{i=1}^{m} N_{reported}(i,j)} \times 100\% \tag{12}$$

$$DR = \frac{\sum_{j=1}^{n}\sum_{i=1}^{m} N_{true}(i,j)}{\sum_{j=1}^{n}\sum_{i=1}^{m} N_{known}(i,j)} \times 100\% \tag{13}$$

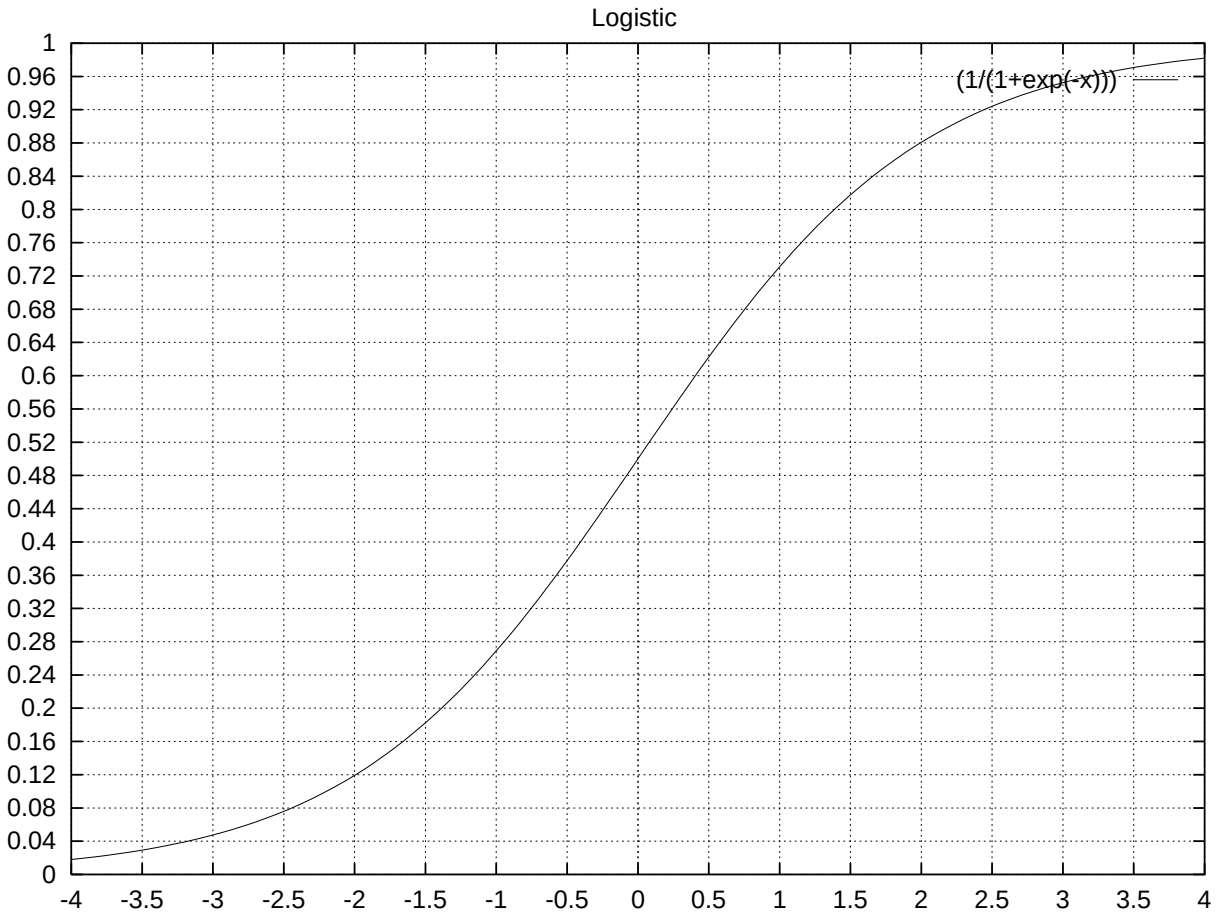
$$FAR = \frac{\sum_{j=1}^{n}\sum_{i=1}^{m} N_{reported}(i,j) - \sum_{j=1}^{n}\sum_{i=1}^{m} N_{true}(i,j)}{\sum_{j=1}^{n}\sum_{i=1}^{m} N_{known}(i,j)} \times 100\% \quad (14)$$

$$P[error(h) > \epsilon] < \delta \quad (15)$$

$$m > (4/\epsilon) ln(4/\delta) \quad (16)$$

$$m \geq max[\frac{1-\epsilon}{\epsilon} ln\frac{1}{\delta}, \frac{VCdim - 1}{32\epsilon}] \quad (17)$$

B Sigmoid/Logistic Function



****