UNIVERSITY OF
STIRLING

INSTITUTE OF COMPUTING
SCIENCE & MATHEMATICS

Computing Science Examination
Spring Semester 2011

CSC9T6: Information Systems

Wednesday 25 May 2011 **09:00-11:00 hours**

Attempt **THREE** questions.

All questions carry equal marks.

The distribution of marks among the parts of each question is indicated.

You may use a nonprogrammable electronic calculator.

IMPORTANT NOTE

It is essential that you write your registration number on the front of each answer book.

Also, when you have completed the examination, the number of answer books that you have used must be prominently written on the front of one book.

1. Imagine you work for a motor insurance company and you want to use data mining techniques to spot fraudulent claims.

   (a) For each of these variables, say whether they are nominal or numeric, discrete or continuous: Age of car, speed of car, type of vehicle (car, bike, lorry). **[4]**

   (b) What form would the output variable take, and what possible values should it take? **[2]**

   (c) How might you collect the data required to build this solution and what challenges might you meet when doing so? **[3]**

   (d) You could use a decision tree or a multi-layer perceptron to model the data. For **each** (that means both!) technique, write a paragraph about the following:

      (i) How its structure represents what is learned and how it makes classifications. Draw a diagram to illustrate your description. **[6]**

      (ii) How it learns – formulae are not required but detail the process of using data to build the model. **[6]**

   (e) What might you need to do to the data before you are able to use it to train a model? Describe two potential problems with data and how you might solve them. **[4]**

2.

(a) Explain what the k-means algorithm is used for (in terms of the data, not an example application). **[2]**

(b) Describe in detail how the k-means algorithm works. As part of your description draw a suitable two dimensional scatter plot showing some example data and the final location of the k-means. **[8]**

(c) Describe a limitation of the k-means algorithm and draw a similar scatter plot to the one you made in part (b) above to illustrate the problem. **[3]**

(d) Describe what it means for a time-series to have each of the following components:

(i) Stationarity. **[2]**

(ii) Cycles. **[2]**

(iii) Seasonality. **[2]**

(e) Describe what the ARMA algorithm is used for and say what the acronym ARMA stands for. **[3]**

(f) Refer to the components in part (d) above and for each, comment on ARMA's ability to model the component. **[3]**

Continued/

3. (a) I travel by train most days. Scotrail use two liveries on their trains: 30% of the time the trains are painted maroon on the outside, and 70% of the time they are painted blue. Further, 20% of the time, the train has a buffet car, 30% of the time there is a refreshment trolley, and 50% there is neither. Assume that the colour of train is independent of whether or not there is a buffet or a trolley.

Given the above, when I go to the train station today after the examination, what is the probability I would find:

(i) A train with maroon livery. **[2]**

(ii) A train with blue livery AND a buffet car. **[2]**

(iii) A train with maroon livery OR a refreshment trolley. **[2]**

(b) Trains do not always run to time. There may be a range of problems causing trains to run late: there may be signalling problems on the line, or a crew member failed to turn up for work. If there are no signalling problems and all crew report for work, then the trains run late 10% of the time. If there are signalling problems (and no crew problem) then 30% of the time the train will run late. If a crew member fails to report for work (but the signals are ok) then 65% of the time the train runs late. If there are signalling problems *and* a crew member fails to report for work, the trains run late 95% of the time. 10% of the time there are signalling problems on the line. 20% of the time a crew member fails to report for work.

(i) Show how this information would be represented as a Bayesian Belief Network. Indicate clearly which node or nodes are parent nodes and which node or nodes are child nodes. Choose suitable names for the nodes of your network. **[5]**

(ii) Bayesian Belief Networks make extensive use of conditional probability information. Explain what is meant by conditional probability, giving an example from the scenario above with its value. **[3]**

(iii) Given the above, calculate the probability of any given train running late. Show your formula for working this out, as well as showing the numeric values of the probabilities. **[5]**

(iv) Given your answer to part (iii), what is the probability of any given train running on time? **[1]**

(v) Given that a train is late, use Bayes' Theorem to calculate the probability that a member of that train crew did not turn up for work (whether or not there are signalling problems). As above, show your formula for working this out, as well as showing the numeric values of the probabilities. **[5]**

Continued/

4. (a) Stirling University wishes to build an automated decision support system to advise students about their academic progress. Students will interact with the system to answer questions such as: What modules do I have to take next semester? What degree programs am I eligible for? What are my options if I fail module X? I want to switch to degree program Y – do I need to take any additional modules? The developer of the system will have access to three sources of knowledge: the University's **Student Record database**, which has information about past and present students, including the modules taken, grades achieved, degree awarded, changes of degree program, and so on; the **University Calendar**, which documents the rules and requirements for all degree programmes; and the **staff** of the Student Programmes office, who are familiar with the University's rules and experienced at advising students.

You have been asked to advise the University on how to implement this system. Two methods are being considered:

a rule-based expert system

a case-based reasoning system.

(i) Explain the principles underlying each of these methods. Describe in detail what would be involved in using each method to implement the student advisor system for Stirling University. **[6]**

(ii) Compare and contrast the two methods by considering these factors: ease of implementation; ease of use; ease of maintenance; trust in conclusions reached by the system. **[4]**

(iii) Which method would you advise the University to use, and why? **[2]**

(b) A fuzzy logic controller is to be used to maintain a comfortable environment inside a classroom by adjusting the speed of a heating fan. When the fan is switched on, it blows warm, dry air into the room, raising the temperature and reducing the relative humidity. The controller receives inputs from a thermometer and a humidity sensor located in the room. The controller uses the following rules:

```
IF temp IS cold AND humidity IS high THEN fan_speed IS
high

IF temp IS cold AND (humidity IS medium OR humidity IS
low)
      THEN fan_speed IS medium

IF temp IS cold AND humidity IS low THEN fan_speed IS low

IF temp IS good AND humidity IS high
      THEN fan_speed IS medium

IF temp IS good AND (humidity IS medium OR humidity IS
low)
```

Continued/

```
      THEN fan_speed IS low

IF temp IS hot THEN fan_speed IS zero
```

Continued/

(i) Draw fully-labelled, appropriately-shaped graphs for the nine fuzzy sets involved. Assume that temperature is expressed in °C, relative humidity is expressed as a percentage, and fan speed is measured in m^3/s (cubic metres per second). Assume that a "good" temperature is around 21°C, "medium" relative humidity is around 45%, and that fan speeds range from 0--50 m^3/s. Explain any further assumptions that you make and justify the shapes you have chosen for your graphs. **[6]**

(ii) Show how the appropriate fan speed is calculated if the sensor inputs have fuzzy values as follows:

Fuzzy Set	**Value**
`temp IS cold`	1.0
`humidity IS medium`	0.6
`humidity IS low`	0.5
all other sets	0

Your answer should include an explanation of the "defuzzification" step, though exact mathematical formulae are not required. **[7]**

END OF EXAMINATION