This question paper consists of 9 printed pages, each of which is identified by the Code Number COMP5830M01.

A non-programmable calculator may be used.
Answer All Questions.
Open Book.
Course notes are permitted.

School of Computing

January 2015

COMP5830M01

KR+ML: KNOWLEDGE REPRESENTATION AND MACHINE LEARNING (MSc)

Time allowed: 2 hours and 15 minutes

PLEASE DO NOT REMOVE THIS PAPER FROM THE EXAM ROOM

Answer ALL FOUR questions

The marks available for each part of each question are clearly indicated.

Question 1

a) Translate the following sentences into *First-Order Predicate Logic* (using equality where necessary):

i) Anyone who likes Tom also likes Joe. [2 marks]

ii) I know a man whose sister knows a friend of Susan [3 marks]

b) Using the *Sequent Calculus* (as specified in the module notes), determine whether the following sequent is valid:

$$\neg(R \wedge \neg S),\ S \rightarrow T,\ R \ \vdash \ T$$

[6 marks]

c) Give a representation of the following statements in *Propositional Tense Logic*:

- If Mary has not bought a ticket she will not go to the ball. [2 marks]
- I have only drunk sake when in Japan. [2 marks]

[15 marks total]

Question 2

a) A *Prolog* database of food items classifies the items according to whether they are a vegetable, a dairy product, a meat or a starch. It also classifies certain items as being high fat or high sugar foods. A food item can potentially fall under more than one (or possibly none) of the categories.

For example, the database might contain facts such as:

- `vegetable( carrot ).`
- `dairy( cheese ).`
- `meat( chicken ).`
- `starch( rice ).`
- `high_fat( cheese ).`
- `high_sugar( meringue ).`

In order to make use of the food database to identify healthy meals it is required that the following predicates be defined to infer implied properties of food items and meals (which will be represented as lists of food items). Specifically you must specify the following predicates:

i) A predicate `protein/1`, such that `protein(X)`, is true just in case `X` is either a `dairy` item a `meat` item or is the particular food item `tofu`. [2 marks]

ii) A predicate `healthy_food/1`, such that `healthy_food(X)`, is true just in case `X` is neither a `high_fat` nor a `high_sugar` item (you may assume that `X` is a food item without explicitly checking this). [2 marks]

iii) A predicate `healthy_meal/1`, such that `healthy_meal(M)`, is true just in case `M` is a list of three food items, such that the first is a `protein`, the second is a `vegetable` and the third is a `starch` and all three of these items are `healthy_foods` (you may assume that this predicate has been defined even if you have not answered the previous part). [4 marks]

b) An AI program is being written to control a robot that will be used to paint components that will be assembled to make toys and ornaments. The robot can pick up objects, move them from one place to another and spray them with paint of various colours. The program will be implemented using *Situation Calculus* to describe the actions that can occur in this scenario and the effects that they will cause.

Write down a *frame axiom* that expresses the knowledge that: [2 marks]

Picking up an object does not change its colour.

c) Consider the following formulae involving topological relations of the *Region Connection Calculus* (RCC), together with the spatial sum function and the convex hull function, conv. The quantifiers range over non-empty spatial regions. For each of the following formulae, draw a configuration of the regions a and b which satisfies the formula. Label your diagram to indicate which region is which:

i) $\neg\mathsf{EQ}(a, \mathsf{conv}(a)))$ [1 mark]

ii) $\mathsf{TPP}(a, b) \wedge \mathsf{DC}(c, a) \wedge \mathsf{EC}(c, b)$ [2 marks]

iii) $\mathsf{DC}(a, b) \wedge \mathsf{NTPP}(a, \mathsf{conv}(b))$ [2 marks]

[15 marks total]

Question 3

a) Consider the following statement:

> *"The ideal Decision Tree is one which exactly fits the training data and should be used to classify all future test data."*

State whether you agree with this claim, justifying your answer. [2 marks]

b) Consider the following dataset, where X1, X2, X3 are input binary random variables, and Y is a binary output whose value we want to predict:

$\mathbf{D}$	$\mathbf{Y}$	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_1}$
D_1	1	0	1	1
D_2	0	0	0	0
D_3	1	1	1	1
D_4	0	0	1	1
D_5	1	1	1	1

Given the input $[\mathbf{X_1}, \mathbf{X_2}, \mathbf{X_3}] = [1, 0, 0]$, what would a Naïve Bayes classifier predict for Y? You must show how you calculate your answer. [3 marks]

c) A big insurance company wants to review its health insurance policy with regards to high blood pressure (B) according to whether their customers exercise regularly and have a good diet. Evidence shows that regular exercise (E) is key to overall good health and fitness (H). But exercise is not enough; in fact maintaining a healthy diet (D) is key to good health, while a poor diet even when exercising can lead to health problems and high blood pressure.

i) Draw a Bayes Net corresponding to the information above. [1 mark]

ii) The company has collected the following data for its customers:

- 60% of people exercise regularly.
- 30% of people have a poor diet.
- There is a 90% chance of high blood pressure if someone has low fitness and bad health, but only a 10% chance otherwise.
- If someone exercises regularly but has a poor diet, then there is a 60% chance they will have low fitness, but only a 15% chance if they eat healthy.
- If someone doesn't exercise regularly and has a poor diet then there is a 90% chance they will have low fitness and bad health.
- However, the chance of bad health is 60% if they have a good diet even if they don't exercise regularly.

The conditional probability tables for some of this information are given below.

Give the remaining conditional probability table for the variable H. [1 mark]

Θ_E	T	F
	0.6	0.4

Θ_D	T	F
	0.7	0.3

Θ_B	T	F
H = T	0.1	0.9
H = F	0.9	0.1

iii) Calculate the probability that someone who exercises regularly and maintains a healthy diet will have high blood pressure. Show your calculations clearly in the answer booklet. [4 marks]

d) Consider the problem of a GP trying to diagnose whether a patient has Tuberculosis based on some common symptoms of it, such as: whether the patient coughs for more than two weeks, has a fever, feels exhausted for no reason, has anorexia and weight loss and feels chest pain. The GP has recorded the following 15 past cases:

Patient	cough duration $>$ 2 weeks	fever	fatigue	anorexia & weight loss	chest pain	tuberculosis
P1	no	low	yes	yes	no	no
P2	no	high	yes	no	yes	no
P3	yes	low	yes	yes	yes	yes
P4	yes	high	yes	yes	yes	yes
P5	yes	high	no	no	yes	yes
P6	yes	none	no	no	yes	no
P7	yes	low	no	no	yes	yes
P8	no	none	yes	no	no	no
P9	yes	low	no	no	no	yes
P10	yes	high	yes	yes	no	yes
P11	yes	high	no	yes	yes	yes
P12	yes	low	yes	no	yes	yes
P13	no	high	no	no	yes	yes
P14	yes	high	yes	no	no	no
P15	yes	low	yes	yes	no	no

i) Assist the GP by constructing an expert system based on a ID3 Decision Tree that describes the 15 past cases. Which variable you would use for the initial split at the root of the tree (hint: it is one of cough or fever). Show how you would decide which one of the two it is, and justify your answer. [3 marks]

Notes: To answer this question, you do not need to draw the whole tree, but only to compute which attribute to split on at the root of the tree.

*If your calculator does not have a function for log2 then you may use the approximation log2(n) = loge(n) * 1.4427 or log2(n) = log10(n) * 3.322.*

ii) Consider the following training instance, P16, added to the above ones, in which a nurse did not notice that the thermometer was broken. Compute again the information gain for Fever. The new gain for Cough is G(Cough)=0.160.

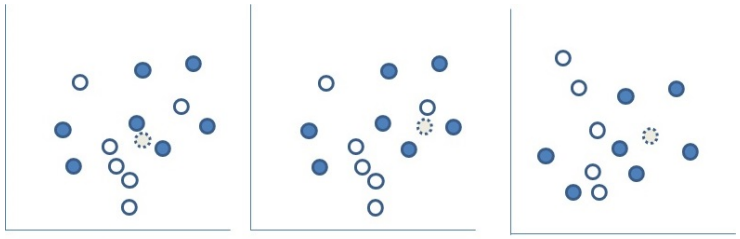
What do you observe? Comment on your findings and state which variable you would now choose for the root node. [1 mark]

Patient	cough duration $>$ 2 weeks	fever	fatigue	anorexia & weight loss	chest pain	tuberculosis
P16	yes	none	no	no	yes	yes

[15 marks total]

Question 4

a) Consider the figure below; find a value of k in a k-nearest neighbourhood classifier, which would classify each of the dotted circles as grey? Justify your answer. How can k be chosen in general? [3 marks]



b) Suppose that at a certain stage in processing the Version Space Candidate Elimination Algorithm has the following version space:

`G-set: [[a,?,?]]`
`S-set: [[a,a,a]]`

Assume that the only values possible for each attribute are `a`, `b`, `c` (for each of the three attributes).

Show the new version space which would result from the above version space in the case that each of the following examples is the next example: [3 marks]

i) positive new example `[a,a,b]`.

ii) negative new example `[a,b,c]`.

iii) positive new example `[b,a,a]`.

c) Consider the following results from a machine learning algorithm:

Test case	Actual class	Predicted class
1	c	c
2	b	b
3	b	b
4	c	b
5	c	c
6	a	a
7	b	c
8	a	a

The *Predicted class* column shows the result from the machine learning algorithm, whilst the *Actual class* column shows the true class for the test case.

Draw a confusion matrix for the above data, making clear what the rows and columns denote. Which classes, if any, are confused? [2 marks]

d) Consider the following results from a machine learning algorithm:

Test case	Actual class	Predicted class
1	1	0
2	1	1
3	1	0
4	1	0
5	0	1
6	0	1
7	0	0
8	0	0

Compute the following measures: [4 marks]

i) the number of true positives (TP)
ii) the number of false positives (FP)
iii) the number of true negatives (TN)
iv) the number of false negatives (FN)
v) the accuracy of the algorithm
vi) the recall (or true positive rate) of the algorithm
vii) the precision of the algorithm
viii) the F1 score of the algorithm

e) How are sample training sets chosen in the bagging machine learning method and how are they used? [3 marks]

[15 marks total]

END