This question paper consists of 14 printed pages, each of which is identified by the Code Number COMP5830M01

******* VERSION WITH ANSWERS INCLUDED *******

School of Computing

January 2014

KRR: KNOWLEDGE REPRESENTATION AND MACHINE LEARNING (MSc)

Time allowed: 2 hours and 15 minutes

Answer ALL FOUR questions.

This is an open book examination. Candidates may take with them into the examination room any written or printed material. Reproduction or simple rephrasing of course notes is unlikely to win credit in any question.

Turn over for question 1

Question 1

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

i) Any teacher who teaches his/her own child is unwise. [2 marks]

Answer:

$\forall x \forall y[(\mathsf{Teacher}(x) \wedge \mathsf{HasChild}(x,y) \wedge \mathsf{Teaches}(x,y)) \rightarrow \mathsf{Unwise}(x)$

ii) Nobody trusts a politician with no children. [2 marks]

Answer:

$\forall x[(\mathsf{Politician}(x) \wedge \neg\exists y[\mathsf{HasChild}(x,y)]) \rightarrow \neg\exists z[\mathsf{Trusts}(z,x)]]$

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

$$\forall x[A(x) \rightarrow (B \wedge C)],\ \neg C \ \vdash\ \forall x[\neg A(x)]$$

[7 marks]

Answer:

The sequent is valid, as shown by the following proof:

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{Axiom}{\forall x[A(x) \rightarrow (B \wedge C)],\ \neg A(k) \ \vdash\ C,\ \neg A(k)} \quad and \quad \dfrac{\dfrac{Axiom}{\forall x[A(x) \rightarrow (B \wedge C)],\ B,\ C \ \vdash\ C,\ \neg A(k)}}{\forall x[A(x) \rightarrow (B \wedge C)],\ (B \wedge C) \ \vdash\ C,\ \neg A(k)}}{\forall x[A(x) \rightarrow (B \wedge C)],\ \neg A(k) \vee (B \wedge C) \ \vdash\ C,\ \neg A(k)}}{\forall x[A(x) \rightarrow (B \wedge C)],\ A(k) \rightarrow (B \wedge C) \ \vdash\ C,\ \neg A(k)}}{\forall x[A(x) \rightarrow (B \wedge C)] \ \vdash\ C,\ \neg A(k)}}{\forall x[A(x) \rightarrow (B \wedge C)],\ \neg C \ \vdash\ \neg A(k)}}{\forall x[A(x) \rightarrow (B \wedge C)],\ \neg C \ \vdash\ \forall x[\neg A(x)]}$$

1 mark for each correct rule application. For full marks complete proof is required with Axioms at the top. Some credit may be given for almost correct rule applications.

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

- In future I shall only drink wine when on holiday. [2 marks]

  Answer:

  $\mathbf{G}(\mathsf{DrinkWine} \rightarrow \mathsf{OnHoliday})$

- If the robot has cleaned the room it will shut down unless it is malfunctioning. [2 marks]

  Answer:

  $(\mathbf{P}\mathsf{RobCleanedRoom} \wedge \neg\mathsf{RobMalfunction}) \rightarrow \mathbf{F}\mathsf{RobShutDown}$

[15 marks total]

Question 2

a) A Prolog program is used to make inferences regarding the students on university courses and the times of lectures on those courses. The program incorporates facts asserted using a `studies/2` predicate that associates each student with his/her course, and a `lecture/4` predicate that relates a course, a room, a time period and a day of the week (meaning that the course includes a lecture in that room at the given time, on the given day).

Examples, of the facts asserted are as follows:

```
studies( john,  maths ).
studies( karen, maths ).
studies( lucy,  physics ).
studies( mark,  computing ).

lecture( maths,   room1, 10-11, monday ).
lecture( maths,   room1, 10-12, tuesday ).
lecture( physics, room2, 10-11, monday ).
lecture( physics, room1, 11-13, tuesday ).
lecture( physics, room1, 12-13, tuesday ).
```

The courses are such that every student on a given course attends every lecture that is part of that course (e.g. John takes all maths lectures).

Time periods of lectures are represented by a term of the form `T1-T2`, with `T1` being the start time and `T2` being the end time. The times are given using the 24 hour clock representation and are always in whole hours (no minutes). You may assume that all time periods lie within a single day, so that the start time is always strictly less than the end time.

Define the following Prolog predicates, satisfying the given definitions of the conditions under which they are true:

i) `same_course( S1, S2 )` — Holds when students `S1` and `S2` are studying on the same course. [1 mark]

ii) `visits_room( S, R )` — Holds when student `S` has at least one lecture in room `R`. [2 marks]

iii) `periods_clash( P1, P2 )` — Holds when the given time periods are such that there is at least one hour which is both within `P1` and within `P2`. [2 marks]

iv) `course_has_clash( C )` [2 marks]

Answer:
The required definitions should be similar to the following:

```
same_course( X, Y ) :- studies( X, S ), studies( Y, S ).

visits_room( X, R ) :- studies( X, S ),
                       lecture( S, R, _, _ ).

periods_clash( B1-E1, B2-E2 ) :- B1 < E2, B2 < E1.
```

```
course_has_clash( C ) :- lecture( C, _, P1, D ),
                         lecture( C, _, P2, D ),
                         \+( P1 = P2),
                         periods_clash( P1, P2 ).
```

b) An AI specialist wants to model a situation in which a group of agents each possess various items and must exchange these with each other in order to acquire a set of items that can be used to build some desired structure. He/she decides to model the scenario using *Situation Calculus*.

Among the fluents used in the representation will be:

- $\mathsf{has}(ag, i)$ — agent ag has item i.
- $\mathsf{wants\text{-}to\text{-}build}(ag, st)$ — agent ag wants to build structure st.

The following (non-fluent) domain predicates will also be used):

- $\mathsf{component}(i, st)$ — item i is a required component of structure st.
- $\mathsf{similar\text{-}value}(i1, i2)$ — items $i1$ and $i2$ have similar value.

In the scenario being modelled, agents will only trade to acquire items that they need to build the structure that they want to build. Moreover, they will only trade an item for another item that is similar in value.

Give a suitable *pre-condition* axiom for the **trade** action. [4 marks]

Answer:

$$Poss(\ \mathbf{trade}(ag1, i1, ag2, i2),\ \ s\ ) \leftarrow$$
$$Holds(\mathsf{has}(ag1, i1), s) \wedge Holds(\mathsf{has}(ag2, i2), s) \wedge \mathsf{similar\text{-}value}(i1, i2) \wedge$$
$$Holds(\mathsf{wants\text{-}to\text{-}build}(ag1, st1), s) \wedge Holds(\mathsf{wants\text{-}to\text{-}build}(ag2, st2), s) \wedge$$
$$\mathsf{component}(i2, st1) \wedge \mathsf{component}(i1, st2)$$

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

ii) $\exists x[\mathsf{NTPP}(a, x) \wedge \mathsf{P}(x, b)]$ [2 marks]

Answer:
Some possibilities are shown below; these are not necessarily unique; others accepted as appropriate.

```
 i)   -------------------         lose a mark for each condition
      |                 |         not satisfied
      |     -------------
      |     |
      | a   |   ---------------------
      |     |   |                   |
      |     |   ---------------|    |
      |     |                  |    |
      |     ---------------    | b  |
      |                   |    |    |
      ---------------------    |    |
                               |    |
                    -----------|    |
                    |               |
                    -----------------

ii)
      -------------------        lose a mark if a touches edge of b
     |  b               |
     |                  |
     |    ----------    |
     |    | a      |    |
     |    |        |    |
     |    ----------    |
     |                  |
     --------------------
```

[15 marks total]

Question 3

a) Consider the following statement: "If a learned decision tree overfits the training data then there is nothing that can be done about it". Explain what this sentence means and say whether you agree with it and why. [2 marks]

Answer:
Overfitting means that the model has not generalised the data sufficiently to predict well on future data. Pruning can be used to reduce the likelihood of overfitting by reducing the specificity of the model (forcing generalisation), e.g. by dropping levels from the tree, or by converting the tree to rules and dropping conjuncts. Cross validation can be used to determine whether pruning is required and how much pruning should take place.

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

	U	X	Y	Z
T1	1	1	0	0
T2	1	0	1	1
T3	0	1	0	0
T4	0	1	1	1
T5	0	0	1	0

Given the input X= 1, Y = 1, Z= 0, what value would a Naïve Bayes classifier predict for U? [3 marks]

Answer:

p(U = 1) = 2/5; p(U = 0) = 3/5

p(X = 1| U = 1) = 1/2; p(X = 1| U= 0) = 2/3; p(Y = 1| U = 1) = 1/2

p(Y = 1| U = 0) = 2/3; p(Z= 0| U = 1) = 1/2; p(Z = 0| U = 0) = 2/3

Predicted U maximizes p(X= 1| U) p(Y = 1| U) p(Z = 0| U) p(U)

p(X = 1| U = 1) p(Y = 1| U = 1) p(Z= 0| U = 1) p(U = 1) =

1/2 * 1/3 * 2/3 * 2/5 = 4/90 = 0.05

p(X = 1| U = 0) p(Y = 1| U = 0) p(Z = 0| U = 0) p(U = 0) =

2/3 * 2/3 * 2/3 * 3/5 = 24/135 = 0.1778

Hence, the predicted U is 0.

[1.5 marks for p(X = 1| U = 1) and 1.5 for p(X = 1| U = 0)]

c) Whether a person brushes their teeth regularly or not (B) affects whether they will have good dental health (H) or not. Eating lots of sweets (S) is bad for dental health. Someone with bad dental health is more likely to need a tooth extracted (E).

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

Answer:

```
B       S
 \     /
  v   v

    H

    |
    v
    E
```

ii) Assume you have the following information:

- 80% of people brush their teeth regularly
- 40% of people eat lots of sweets
- There is an 25% chance of requiring an extraction if you have bad dental health, but only a 1% chance otherwise.
- If someone brushes their teeth regularly and eats lot of sweets, then there is a 45% chance they will have bad dental health, but only a 2% chance if they don't eat lots of sweets.
- If someone doesn't brush their teeth regularly and eats lot of sweets, then there is a 99% chance they will have bad dental health, but only a 90% chance if they don't eat lots of sweets.

The conditional probability tables for some of this information are given below. Give the remaining conditional probability table for the variable H. [2 marks]

θ_B	T	F
	0.8	0.2

θ_S	T	F
	0.4	0.6

θ_E	T	F
H=T	0.01	0.99
H=F	0.25	0.75

Answer:

θ_H	B	S	T	F
	T	T	.55	0.45
	T	F	.98	0.02
	F	T	0.01	.99
	F	F	0.1	.9

Marks: half a mark per row correct

iii) Calculate the probability that someone who brushes their teeth regularly and doesn't eat lots of sweets will need an extraction.

(*Hint:* You should get an expression for p(B,S,E) by marginalising over H, and another by using the product rule and conditional independence. Combining these should allow you to express p(E|B,S) in terms of the variables B, S, H and E.) [4 marks]

Answer:

The question asks p(E=T | B=T, S=F) = ?

By the product rule and conditional independence

p(E,B,S) = p(E| B,S) p(B,S)

= p(E| B,S) p(B) p(S)

Using the sum rule to marginalise over H:

p(B,S,H) = $\Sigma_{x \in H}$ p(B,S,H=x,E)

From the graph p(B,S,H,E) = p(E| H) p(H| B,S) p(B) p(S)

Calculating the above:

p(E| B,S) = p(B,S,E) / [p(E) p(S)]

= Σ_H p(B,S,H,E)/ p(B) p(V)

= [Σ_H p(E| H) p(H| B,S) p(B) p(S)] / p(B) p(S)

= Σ_H p(E|H) p(H| B,S)

So p(E=T | B=T, S=F) = p(E=T | H=T) p(H=T | B=T, S=F) + p(E=T | H = F) p(H=F| B=T, S=F)

From the tables:

= 0.01 * 0.98 + 0.25 * 0.02

= .0098 + .005

= 0.0148

[1 mark for correct answer; 3 for calculations]

d) Consider the following table which lists 10 training examples with attributes x, y, and z where w is the target class.

	x	y	z	w
D1	1	1	1	n
D2	0	1	2	y
D3	0	1	3	y
D4	0	1	4	y
D5	1	0	8	n
D6	0	0	9	n
D7	0	1	1	y
D8	1	0	5	y
D9	1	1	6	n
D10	1	1	7	n

Consider the task of constructing a Decision Tree for this set of training instances and whether to split on x or y (the variable z has many values so is not deemed a good initial choice). Suppose you already know that the information gain for splitting on the attribute x would be 0.278072.

Compute the information gain if the attribute y were to be chosen and hence whether it is better to split on x or on y. You may use the fact that the expected number of bits $I(e)$ to describe a distribution given evidence e is $\Sigma_x - p(x \mid e) * \log_2 p(x \mid e)$ and the information gain from a test α is $I(true) - (p(\alpha) * I(\alpha) + p(\neg\alpha) * I(\neg\alpha))$. *Note: you do not need to draw the whole tree, only to compute which attribute to split on at the root of the tree.*

(*If your calculator does not have a function for* $\log_2$ *then you may use the approximation* $\log_2(n) = \log_e(n) * 1.4427$ *or* $\log_2(n) = \log_{10}(n) * 3.322$.)

[3 marks]

Answer:

First compute the entropy of the overall data set E (5+,5-), so entropy is 0.5*log(0.5) + 0.5 * log(0.5) = 0.5 + 0.5 = 1

Suppose split on attribute “y”:

y = 1 we get (4+; 3-). Entropy is 4/7 * log(4/7) - 3/7 log(3/7) = 0.985228136

y = 0 we get (1+; 2-). Entropy is 1/3 * log(1/3) - 2/3 log(2/3) = 0.918295834

The information gain for attribute “y” is computed by

1 - (0.7 * 0.985228136 + 0.3 * 0.918295834) = 0.034851555

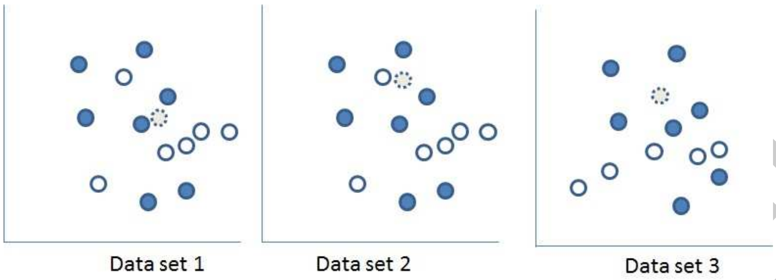
More gain for x so split on x.

[1 mark for overall entropy calculation; 2 for calculations on v.]

[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? How can k be chosen in general? [2 marks]



Answer:

k=3. [1.5 marks]

k can be estimated heuristically, e.g. using cross validation. [0.5 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: [4 marks]

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

Answer:

i) G-set: `[[?,?,a]]` S-set: `[[a,?,a]]`
ii) G-set: `[[a,?,a],[?,a,a]` S-set: `[[a,a,a]]`
iii) There is no consistent version space

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

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

Answer:

Confusion matrix

class	a	b	c
a	1	0	1
b	1	3	0
c	0	0	2

The rows are the actual classes and the columns the predicted classes. [1 mark]

The classes a and b are confused whilst c is perfectly classified. [1 mark]

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

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

Answer:

Test case	Actual class	Predicted class	TP	TN	FP	FN
1	1	1	1	0	0	0
2	1	1	1	0	0	0
3	1	0	0	0	0	1
4	1	0	0	0	0	1
5	0	1	0	0	1	0
6	0	1	0	0	1	0
7	0	1	0	0	1	0
8	0	0	0	1	0	0
			2	1	3	2
P	4					
N	4					
Accuracy	0.375					
Precision	0.4					
Recall	0.5					
F1	0.444					

[0.5 mark per measure]

e) What is meant by “resampling with replacement”? Briefly describe one machine learning technique which uses “resampling with replacement” and explain its advantages. [3 marks]

Answer:
“Resampling with replacement” means that after sampling a training set for a learning a model the items are put back in so that the next sample could choose some of the same ones [1]. Bagging is a bootstrap (ensemble) method applied to learning ensemble of classifiers. Each base learner is trained with a resampled training set. All the base learners vote with the same weight. It reduces variance and helps to avoid overfitting – e.g. voting can amount to taking the mean, which will help smooth overfitted models – though not useful for linear models [2].

[15 marks total]

END