This question paper consists of 16 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]

Answer:

$\forall x[\mathsf{Likes}(x, \mathsf{tom}) \rightarrow \mathsf{Likes}(x, \mathsf{joe})]$.

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

$\exists x[\mathsf{Man}(x) \wedge \mathsf{Knows}(\mathsf{me}, x) \wedge \exists y \exists z[\mathsf{Sister}(x, y) \wedge \mathsf{Friend}(z, \mathsf{Susan}) \wedge \mathsf{Knows}(y, z)]]$

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]

Answer:

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

```
    Axiom                    Axiom
--------------          ---------------
-S, R  |-  T, R          -S, R  |- T, -S
------------------------------------ |-&
      -S, R  |-  T, (R & -S)                T, R  |-  T, (R & -S)
     ------------------------------------------------------------ v |-
                    -S v T, R  |-  T, (R & -S)
                    ------------------------- -> r.w.
                    S -> T, R  |-  T, (R & -S)
                    -------------------------  - |-
                    -(R & -S), S -> T, R  |-  T
```

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*:

- If Mary has not bought a ticket she will not go to the ball. [2 marks]
  Answer:
  $\neg\mathbf{P}\mathsf{MBT} \rightarrow \neg\mathbf{F}\mathsf{MGB}$
- I have only drunk sake when in Japan. [2 marks]
  Answer:
  $\mathbf{H}(\mathsf{IdrinkSake} \rightarrow \mathsf{IinJapan})$

[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]

Answer:

```
protein( P ) :- meat( P ) ; dairy( P ) ; P = tofu.                    2 marks

healthy_food(F) :- \+high_sugar(F),                                    2 marks
                   \+high_fat(F).

healthy_meal( [P,V,S] ) :- protein(P),
                           vegetable( V ),                             4 marks
                           starch(S),
                           healthy_food(P), healthy_foot(V), healthy_food(S).
```

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.

Answer:

$\forall xcs[Holds(\text{colour}(x, c), s) \rightarrow Holds(\text{colour}(x, c), result(\mathbf{pickup}(x), s))]$

Some variants may also be correct. The leading universal quantifier may be omitted.

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:

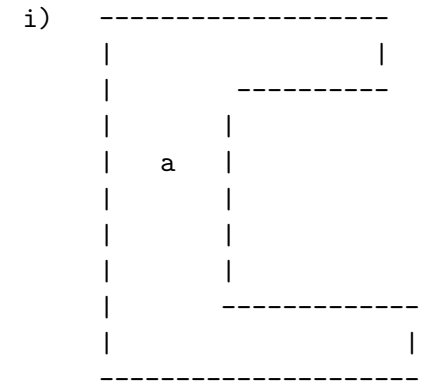
i) $\neg\text{EQ}(a, \text{conv}(a)))$ [1 mark]

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

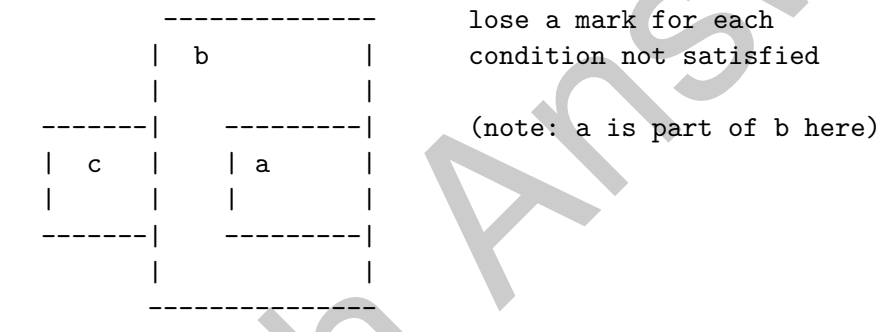
iii) $\text{DC}(a, b) \wedge \text{NTPP}(a, \text{conv}(b))$ [2 marks]

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

i)



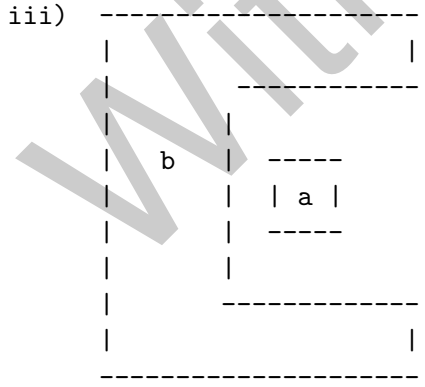
ii)



lose a mark for each condition not satisfied

(note: a is part of b here)

iii)



lose a mark for each condition not satisfied

[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]

Answer:
I would agree that the ideal decision tree is the one that perfectly fits the training data if we are certain that this training dataset is noiseless. In practice and in most real world situations the assumption of perfect training data rarely holds true. As such, a DT fitting perfectly a training dataset would also fit perfectly the noisy training instances, resulting in an illusion of good performance when in reality the actual accuracy is lower, as it tends to be discovered later when trying to classify new unseen instances. This problem is commonly called overfitting and it is a problem with a variety of ML methods. In the case of the decision trees we can use different techniques to avoid overfitting. These are based on pruning the DT, and they are either post-pruning pre-pruning. Post-pruning methods are more common and easier than pre-pruning. Post-pruning methods are cross-validation (reduced error pruning) techniques where the training set is split in training and validation datasets, then the whole tree is expanded to its best solution and then branches are removed greedily as long as they minimise the accuracy error. Most efficient is rule post pruning in which the tree is transformed in conditional statements. Then the conditions of each statement is removed if the validation accuracy is improved. In the end the rules are sorted according to their accuracy values. Pre-pruning methods do not allow the tree to grow during its training if there is no statistical significant improvement (error estimation of subtree, chi-squared test between accuracy of sample and complete distribution). There are also other techniques such minimal descriptor length that try to minimise both the length of the tree as well as the information transmitted which is the misclassifications.

[1mark for saying the bit that disagree as there might be noise in the training data (0.5 marks) and this is commonly known as overfitting (0.5 marks). Overfitting in DTs can be avoided by post-pruning and pre-pruning (0.5 marks). The rest of the details (0.5 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]

Answer:

1 mark for:
p(Y=0) = 2/5
p(Y=1) = 3/5

p(X1=1|Y=0) = 0/2
p(X1=1|Y=1) = 2/3

p(X2=0|Y=0) = 1/2
p(X2=0|Y=1) = 0/3

p(X3=0|Y=0) = 1/2
p(X3=0|Y=1) = 1/3

1 mark for:
Predicted Y maximizes: p(X1=1|Y) p(X2=0|Y) p(X3=0|Y) p(Y)

For Y=0:
p(X1=1|Y=0) p(X2=0|Y=0) p(X3=0|Y=0) p(Y=0) = 0/2 * 1/2 * 1/2 * 2/5 = 0.1

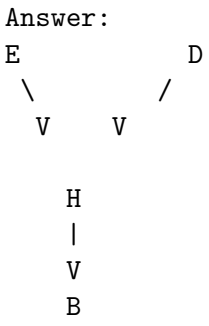
1 mark for:
For Y=1:
p(X1=1|Y=1) p(X2=0|Y=1) p(X3=0|Y=1) p(Y=1) = 2/3 * 0/3 * 1/3 * 3/5 = 0.133

Hence the predicted Y is 1.

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]

Answer:



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

Answer:

Θ_H	E	D	T	F	
	T	T	0.85	0.15	
	T	T	0.4	0.6	0.25 marks per correct row
	T	T	0.6	0.4	
	F	F	0.1	0.9	

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]

Answer:

```
The question asks p(B=T | E=T, D=T) = ?

By the product rule and conditional independence:
p(B, E, D) = p(B|E, D) p(E, D) = p(B|E, D) p(E) p(D)

Using the sum rule to marginalize over H:
p(E, D, H) = Sum{x in H} p(E, D, H=x, B)
From the graph p(E, D, H, B) = p(B|H) p(H|E, D) p(E) p(D)

Calculating the above:
p(B|E, D) = p(E, D, B) / [p(B) p(D)]
        = Sum{over H} p(E, D, H, B) / p(E) p(V)
        = [Sum{over H} p(B|H) p(H|E, D) p(E) p(D)] / p(E) p(D)
        = Sum{over H} p(B|H) p(H|E, D)

So  p(B=T | E=T, D=F) =
= p(B=T|H=T) p(H=T|E=T, D=T) + p(B=T|H=F) p(H=F|E=T, D=T)
From the tables:
= 0.1*0.85 + 0.9*0.15
= .22
```

So the probability that someone who exercises regularly and maintains a healthy diet will have high blood pressure is 22% (stress, genes, environment, smoking, high alcohol consumption, etc. can be other reasons). [1 mark for correct answer; 3 for calculations]

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.*

Answer:

```
We are going to decide on which variable to use as the root node
based on the information gain that each variable provides.

Cough=
Yes(11): 8+ 3-; H = -8/11 log2(8/11)  3/11 log2(3/11) = 0.845
No(4): 1+ 3-; H = -3/11 log2(3/11)  3/11 log2(3/11) = 0.811
G(cough) = 0.135

Fever=
Low(6): 4+ 2-; H = -4/6 log2(4/6)  2/6 log2(2/6) = 0.918
High(7): 5+ 2-; H = -5/7 log2(5/7) 2/7 log2(2/7) = 0.863
None(2): 0+ 2-; H = 0.0
G(fever) = 0.201

% Chest=
```

```
% Yes(9): 7+ 2-; H = 0.764
% No(6): 2+ 4-; H = 0.918
% G(Chest) = 0.145

Fever is the variable to split on as it has the highest information gain.

[1 mark for saying that they will use information gain, 1 mark for
  showing how to compute the information gain, 1 mark if they get all
  answers correct and decide to split on Fever (0.25 for each right
  number)]
```

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

Answer:

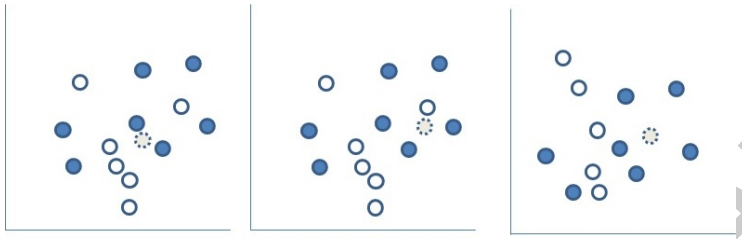
```
Fever=
  Low(6): 4+ 2-, H=0.918 (not marked)
  High(7): 5+ 2-, H = 0.863 (not marked)
  None(3): 1+ 2-, H=0.918
  G_D16(Fever) = 0.971 - (7./16)*0.863 - (6./16)*0.918 - (3./16)*0.918 = 0.077
```

With D16 G(Fever) = 0.077. The best choice to split on is the variable now Chest. The information gain of the variable fever has decreased due to the fact that it no longer best separates the data. In fact, the new training instance that is a noisy one is 1/3 of the training instances for Fever=None and that particular subset has a higher entropy from before, which in turn contributes to lower information gain than before.

[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]



Answer:

k=3. [1 marks]

In each case, of the three nearest neighbours of the dotted circles, at least two are grey. [1]

k can be estimated heuristically, e.g. using cross validation. Choosing odd values is also good if just counting rather than weighting. [1 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]`.

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	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]

Answer:

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

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

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

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

Answer:

Test case	Actual class	Predicted class	TP	TN	FP	FN
1	1	0	0	0	0	1
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	0	0	1	0	0
8	0	0	0	1	0	0
			1	2	2	3
P	4					
N	4					
Accuracy	0.375					
Precision	0.333333333					
Recall	0.25					
F1	0.286					

[0.5 mark per measure]

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

Answer:

"Resampling with replacement" is the methodology: it 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