Name: Student number:

COMP9417 Machine Learning and Data Mining

Mid-session Examination:

SAMPLE QUESTIONS

Multiple-choice questions require **only one** answer.

Show all working in your script book.

This page intentionally left blank.

Question 1 [Total marks: 6]

Well-posed Machine Learning problems

(a) [3 marks] What is required to define a well-posed learning problem ?

(b) [3 marks] Here are two potential real-world application tasks for machine learning:

1. a winery wishes to uncover relationships between records of the quantitative analyses of its wines from the lab and some key subjective descriptions applied to its wine (e.g. dry, fruity, light, etc.)

2. you want to predict students' marks in the final exam of COMP9417 given their marks from the other assessable components in the course — you may assume that the corresponding data from previous years is available

Pick **one** of the tasks and state how you would define it as a well-posed machine learning problem in terms of the above requirements, and Suggest a learning algorithm for the problem you chose (give the name, and in a sentence explain why it would be a good choice).

Question 2 [Total marks: 14]

Concept Learning

A concept learning task has three attributes: `Size`, `Colour` and `Shape`. The possible values for each attribute are:

Size	Colour	Shape
Large	Red	Circle
Small	Green	Triangle
	Blue	Hexagon

Hypotheses are 3-tuples of constraints on the attributes. More precisely, each hypothesis h in the hypothesis space H is a conjunction of 3 constraints of the form $x =?$ or $x = c$ or $x = \emptyset$, where x is one of the attributes, ? means the attribute can have any value, c means the attribute has one of the specific values shown above, and $\emptyset$ means the attribute is not allowed to take on any value. For example, the hypothesis (?, `Green`, ?) defines the concept of a green shape of any size.

(a) [9 marks] Apply the CANDIDATE ELIMINATION ALGORITHM to the three examples given below *in the order in which they appear*. Be sure to show the G and S sets at initialisation, and following input of each example to the algorithm. [Hint: it may help to use a compact notation for hypotheses.]

Example	Instance	Class
1	(`Large`, `Red`, `Circle`)	Negative
2	(`Small`, `Red`, `Circle`)	Positive
3	(`Small`, `Green`, `Hexagon`)	Negative

(b) [2 marks] After learning from these three examples, how would the Version Space classify the instance (`Small`, `Blue`, `Circle`) ?

(c) [3 marks] Now suppose this instance is classified "Positive" and input to the CANDIDATE ELIMINATION ALGORITHM as the next example. Show the G and S sets computed by the algorithm following this input. Has the algorithm learned the target concept ? If you think it has, state the target concept and give a one sentence explanation of your answer. Otherwise, explain in not more than two sentences why it has not.

Question 3 [Total marks: 18]

Decision Tree Learning

(a) [3 marks] Describe the main steps in the basic decision tree learning algorithm.

The table below contains a sample S of ten examples. Each example is described using two Boolean attributes A and B. Each is labelled (classified) by the target Boolean function.

Id	A	B	Class
1	1	0	+
2	0	1	-
3	1	1	-
4	1	0	+
5	1	1	-
6	1	1	-
7	0	0	+
8	1	1	+
9	0	0	+
10	0	0	-

(b) [2 marks] What is the entropy of thse examples with respect to the given classification ? [Note: you must show how you got your answer using the standard formula.]

This table gives approximate values of entropy for frequencies of positive examples in a two-class sample.

Frequency of class '+' in sample	Entropy of sample
0.0	0.00
0.1	0.47
0.2	0.72
0.3	0.88
0.4	0.97
0.5	1.00
0.6	0.97
0.7	0.88
0.8	0.72
0.9	0.47
1.0	0.00

(c) [**4 marks**] What is the information gain of attribute A on sample S above ?

(d) [**4 marks**] What is the information gain of attribute B on sample S above ?

(e) [**2 marks**] Which would be chosen as the "best" attribute by a decision tree learner using the information gain splitting criterion ? Why ?

(f) [**3 marks**] Describe a method for overfitting-avoidance in decision tree learning.

Question 4 [Total marks: 10]

Classification and Association Rule Learning

Here is a set of examples for a version of the "*EnjoySport*" problem.

Sky	Wind	Water	Outlook	EnjoySport
Sunny	Strong	Warm	Change	Yes
Rainy	Weak	Warm	Same	Yes
Sunny	Weak	Cool	Change	No
Rainy	Weak	Cool	Change	No
Sunny	Strong	Warm	Same	Yes

Trace out by hand the steps in applying the **Decision Table** learning algorithm to the "*EnjoySport*" data above. Use error on the data set rather than cross-validation, and stop as soon as error cannot be improved on the current iteration. In particular, be sure to write down each attribute considered, its evaluation on the training data, and indicate the attribute(s) in the final table that would be selected by the algorithm.

(b) [1 mark] An itemset is called *large* or *frequent* if
(i) the number of items in the itemset exceeds a certain minimum level
(ii) the confidence of the itemset exceeds a certain minimum level
(iii) the number of transactions in the cover of the itemset exceeds a certain minimum level
(iv) none of these

(c) [1 mark] An efficient algorithm for generating candidate k-frequent itemsets can be based on the fact that
(i) no size $k-1$ subset of a k-frequent itemset can be frequent
(ii) no size $k+1$ superset of a k-frequent itemset can be frequent
(iii) all size $k-1$ subsets of a k-frequent itemset are frequent
(iv) all size $k+1$ supersets of a k-frequent itemset are frequent

Question 5 [Total marks: 10]

Learning for Numeric Prediction

(a) Let the weights of a two-input perceptron be: $w_0 = 0.2$, $w_1 = 0.5$ and $w_2 = 0.5$. Assuming that $x_0 = 1$, what is the output of the perceptron when:

[i] **[1 mark]** $x_1 = -1$ and $x_2 = -1$?

[ii] **[1 mark]** $x_1 = -1$ and $x_2 = 1$?

Letting $w_0 = -0.2$ and keeping $x_0 = 1$, $w_1 = 0.5$ and $w_2 = 0.5$, what is the perceptron output when:

[iii] **[1 mark]** $x_1 = 1$ and $x_2 = -1$?

[iv] **[1 mark]** $x_1 = 1$ and $x_2 = 1$?

(b) **[6 marks]** Here is a regression tree with leaf nodes denoted `A`, `B` and `C`:

```
X <= 5 : A
X >  5 :
|   X <= 9: B
|   X >  9: C
```

This is the training set from which the regression tree was learned:

X	Class
1	8
3	11
4	8
6	3
7	6
8	2
9	5
11	12
12	15
14	15

Write down the output (class) values and number of instances that appear in each of the leaf nodes `A`, `B` and `C` of the tree.