Name: Student number:

COMP9417 Machine Learning and Data Mining

SAMPLE: Mid-session Examination

Your **Name** and **Student number** must appear at the head of this page.

Duration of the exam: 1 hour.

This examination has **five** questions. Answer **all** questions.

Total marks available in the exam: 50.

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

Show all working in your script book.

Paper is **NOT** to be retained by the candidate.

This page intentionally left blank.

Question 1 [Total marks: 5]

Well-posed Machine Learning problems

(a) [**1 mark**] 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.

(c) [**1 mark**] 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: 6]

Concept Learning

(a) [**3 marks**] Write an algorithm called "FIND-G" to find a maximally-general consistent hypothesis. You can assume the data will be noise-free and that the target concept is in the hypothesis space.

(b) [**3 marks**] Outline the steps in a proof that FIND-G will never fail to cover a positive example in the training set.

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]

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.

Question 5 [Total marks: 11]

Neural and Tree Learning on Continuous Attributes

(a) [1 mark] In general, feedforward neural networks (multi-layer perceptrons) trained by error back-propagation are:

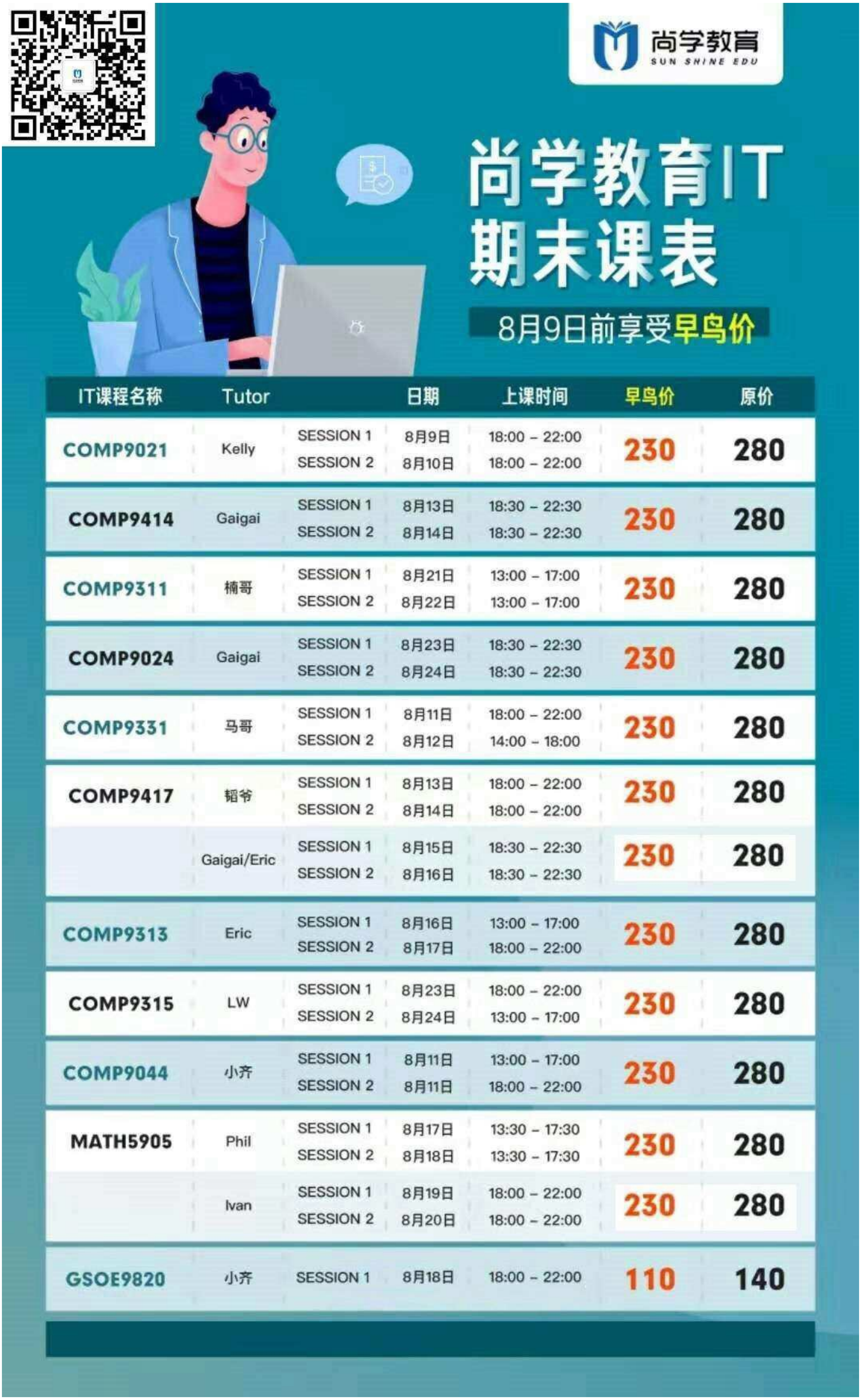
(i) fast to train, and fast to run on unseen examples
(ii) slow to train, and fast to run on unseen examples
(iii) fast to train, and slow to run on unseen examples
(iv) slow to train, and slow to run on unseen examples

In one sentence, explain your choice of answer.

Suppose you have a decision tree (DT) and a multi-layer perceptron (MLP) that have been trained on data sampled from a two-class target function, with all attributes numeric. More generally, you can think of both model classes as graphs, whose edges are labelled with numerical values: *weights* in the MLP and *threshold constants* for feature tests in the DT.

(b) [4 marks] Compare and contrast the *roles* of these numerical values in the two model classes, i.e., for each kind of model, explain how they are used to implement the learned function.

(c) [6 marks] Compare and contrast the *methods of learning* these numerical values in the two model classes, i.e., for each kind of learning algorithm, explain how it will determine these numerical values, given a training set.



尚学教育
SUN SHINE EDU

尚学教育IT
期末课表

8月9日前享受早鸟价

IT课程名称	Tutor		日期	上课时间	早鸟价	原价
COMP9021	Kelly	SESSION 1	8月9日	18:00 – 22:00	230	280
		SESSION 2	8月10日	18:00 – 22:00		
COMP9414	Gaigai	SESSION 1	8月13日	18:30 – 22:30	230	280
		SESSION 2	8月14日	18:30 – 22:30		
COMP9311	楠哥	SESSION 1	8月21日	13:00 – 17:00	230	280
		SESSION 2	8月22日	13:00 – 17:00		
COMP9024	Gaigai	SESSION 1	8月23日	18:30 – 22:30	230	280
		SESSION 2	8月24日	18:30 – 22:30		
COMP9331	马哥	SESSION 1	8月11日	18:00 – 22:00	230	280
		SESSION 2	8月12日	14:00 – 18:00		
COMP9417	韬爷	SESSION 1	8月13日	18:00 – 22:00	230	280
		SESSION 2	8月14日	18:00 – 22:00		
	Gaigai/Eric	SESSION 1	8月15日	18:30 – 22:30	230	280
		SESSION 2	8月16日	18:30 – 22:30		
COMP9313	Eric	SESSION 1	8月16日	13:00 – 17:00	230	280
		SESSION 2	8月17日	18:00 – 22:00		
COMP9315	LW	SESSION 1	8月23日	18:00 – 22:00	230	280
		SESSION 2	8月24日	13:00 – 17:00		
COMP9044	小齐	SESSION 1	8月11日	13:00 – 17:00	230	280
		SESSION 2	8月11日	18:00 – 22:00		
MATH5905	Phil	SESSION 1	8月17日	13:30 – 17:30	230	280
		SESSION 2	8月18日	13:30 – 17:30		
	Ivan	SESSION 1	8月19日	18:00 – 22:00	230	280
		SESSION 2	8月20日	18:00 – 22:00		
GSOE9820	小齐	SESSION 1	8月18日	18:00 – 22:00	110	140