Name of Candidate: ....

Student number: ....

Signature: ....

COMP9417 Machine Learning and Data Mining

Final Examination: SAMPLE QUESTIONS (+ TWO ANSWERS)

Here are six questions which are *somewhat* representative of the type that will be in the final exam. Each question is of equal value, and should not take longer than 30 minutes to complete. In the actual exam there will be some degree of choice as to which questions you can answer. Candidates may bring authorised calculators to the examination, but no other materials will be permitted.

This page intentionally left blank.

Question 1 [20 marks]

Comparing Lazy and Eager Learning

The following truth table gives an "m–of–n function" for three Boolean variables, where "1" denotes true and "0" denotes false. In this case the target function is: "exactly two out of three variables are true".

X	Y	Z	Class
0	0	0	false
0	0	1	false
0	1	0	false
0	1	1	true
1	0	0	false
1	0	1	true
1	1	0	true
1	1	1	false

A) **[4 marks]**
Construct a decision tree which is complete and correct for the examples in the table. [Hint: draw a diagram.]

B) **[4 marks]**
Construct a set of **ordered** classification rules which is complete and correct for the examples in the table. [Hint: use an *if–then–else* representation.]

C) **[10 marks]**
Suppose we define a simple measure of distance between two equal length strings of Boolean values, as follows. The distance between two such strings B_1 and B_2 is:

$$\text{distance}(B_1, B_2) = |(\sum B_1) - (\sum B_2)|$$

where $\sum B_i$ is simply the number of variables with value 1 in string B_i. For example:

$$\text{distance}(\langle 0,0,0\rangle, \langle 1,1,1\rangle) = |0-3| = 3$$

and

$$\text{distance}(\langle 1,0,0\rangle, \langle 0,1,0\rangle) = |1-1| = 0$$

What is the LOOCV ("Leave–one–out cross-validation") error of 2–Nearest Neighbour using our distance function on the examples in the table ? [Show your working.]

D) **[2 marks]**
Compare your three models. Which do you conclude provides a better representation for this particular problem ? Give your reasoning (one sentence).

Question 2 [20 marks]

Bayesian Learning

A) **[4 marks]**
Explain the difference between the *maximum a posteriori* hypothesis H_{MAP} and the *maximum likelihood* hypothesis H_{ML}.

B) **[2 marks]**
Consider a two-class learning problem to "Play tennis", with two Boolean attributes, "Cloudy" and "Windy". Draw the Bayesian network corresponding to a Naive Bayes classifier for this problem.

C) **[10 marks]**
Given the following examples, calculate *all* the probabilities required for your Naive Bayes classifier to be able to decide whether to play or not:

Instance No.	Cloudy	Windy	Play tennis
1	0	0	no
2	0	1	no
3	1	1	no
4	0	0	no
5	0	1	yes
6	1	0	yes

D) **[4 marks]**
To which class would your Naive Bayes classifier assign each of the following instances ?

Instance No.	Cloudy	Windy	Play tennis
7	0	0	?
8	0	1	?

Question 3 [20 marks]

Neural Networks

A) **[4 marks]**
A *linear unit* from neural networks is a linear model for numeric prediction that is fitted by gradient descent. Explain the differences between the *batch* and *incremental* (or *stochastic*) versions of gradient descent.

B) **[4 marks]**
Stochastic gradient descent would be expected to deal better with local minima during learning than batch gradient descent – true or false ? Explain your reasoning.

A) **[12 marks]**
Suppose a single unit has output o of the form:

$$o = w_0 + w_1 x_1 + w_1 x_1^2 + w_2 x_2 + w_2 x_2^2 + \cdots + w_n x_n + w_n x_n^2$$

The problem is to learn a set of weights w_i that minimize squared error. Derive a batch gradient descent training rule for this unit.

Question 3 ANSWER

A) For a linear unit, batch and stochastic gradient descent differ as follows:

- in batch gradient descent the gradient is computed over *all* the examples in the training set before the weight update is applied
- in stochastic gradient descent the gradient is computed for a *single* example, and then the weight update is applied

B) With batch gradient descent, the direction of the gradient is computed over all the training data, so in some sense this is the true gradient. If the algorithm is located near some local minimum then it will move in the direction of the steepest descent and converge at that minimum. However, in stochastic gradient descent, where some example is selected at random, we would expect that the gradient computed for that example may not be the true gradient, so the algorithm may instead move in a different direction, thus avoiding the local minimum. So our answer is 'true'.

C) The approach to deriving the required training rule pretty much follows the method in the lecture slides. We have a single unit with output o of the form:

$$o = w_0 + w_1 x_1 + w_1 x_1^2 + w_2 x_2 + w_2 x_2^2 + \cdots + w_n x_n + w_n x_n^2$$

Using homogeneous coordinates we can write this as:

$$o = \Sigma_{i=0}^{n} w_i (x_i + x_i^2)$$

Assuming the same error function and gradient definition as before (slides 11-14 on the lecture on Neural Networks) we can derive the following:

$$\begin{aligned}
\frac{\partial E}{\partial w_i} &= \frac{\partial}{\partial w_i} \frac{1}{2} \sum_d (t_d - o_d)^2 \\
&= \frac{1}{2} \sum_d \frac{\partial}{\partial w_i} (t_d - o_d)^2 \\
&= \frac{1}{2} \sum_d 2(t_d - o_d) \frac{\partial}{\partial w_i} (t_d - o_d) \\
&= \sum_d (t_d - o_d) \frac{\partial}{\partial w_i} (t_d - \Sigma_{i=0}^{n} w_i (x_i + x_i^2)) \\
\frac{\partial E}{\partial w_i} &= \sum_d (t_d - o_d)(-x_{i,d} - x_{i,d}^2)
\end{aligned}$$

Question 4 [20 marks]

Ensemble Learning

A) **[8 marks]**
As model complexity increases from low to high, what effect does this have on:

1) Bias ?
2) Variance ?
3) Predictive accuracy on training data ?
4) Predictive accuracy on test data ?

B) **[3 marks]**
Is decision tree learning relatively stable ? Describe decision tree learning in terms of bias and variance in no more than two sentences.

C) **[3 marks]**
Is nearest neighbour relatively stable ? Describe nearest neighbour in terms of bias and variance in no more than two sentences.

D) **[3 marks]**
Bagging reduces bias. True or false ? Give a one sentence explanation of your answer.

E) **[3 marks]**
Boosting reduces variance. True or false ? Give a one sentence explanation of your answer.

Question 5 [20 Marks]

Computational Learning Theory

A) **[8 marks]**
An instance space X is defined using m Boolean attributes. Let the hypothesis space H be the set of decision trees defined on X (you can assume two classes). What is the largest set of instances in this setting which is shattered by H ? [Show your reasoning.]

B) **[10 marks]**
Suppose we have a consistent learner with a hypothesis space restricted to conjunctions of exactly 8 attributes, each with values $\{\text{true}, \text{false}, \text{don't care}\}$. What is the size of this learner's hypothesis space ? Give the formula for the number of examples sufficient to learn with probability at least 95% an approximation of any hypothesis in this space with error of at most 10%. [Note: you are *not* required to compute the solution.]

C) **[2 marks]**
Informally, which of the following are consequences of the No Free Lunch theorem:

a) averaged over all possible training sets, the variance of a learning algorithm dominates its bias

b) averaged over all possible training sets, no learning algorithm has a better off-training set error than any other

c) averaged over all possible target concepts, the bias of a learning algorithm dominates its variance

d) averaged over all possible target concepts, no learning algorithm has a better off-training set error than any other

e) averaged over all possible target concepts and training sets, no learning algorithm is independent of the choice of representation in terms of its classification error

Question 6 [20 Marks]

Mistake Bounds

Consider the following learning problem on an instance space which has only one feature, i.e., each instance is a *single integer*. Suppose instances are always in the range $[1, 5]$. The hypothesis space is one in which each hypothesis is an interval over the integers. More precisely, each hypothesis h in the hypothesis space H is an interval of the form $a \leq x \leq b$, where a and b are integer constants and x refers to the instance. For example, the hypothesis $3 \leq x \leq 5$ classifies the integers 3, 4 and 5 as positive and all others as negative.

Instance	**Class**
1	Negative
2	Positive
3	Positive
4	Positive
5	Negative

A) **[15 marks]**
Apply the Halving Algorithm to the five examples in the order in which they appear in the table above. Show each class prediction and whether or not it is a mistake, plus the initial hypothesis set and the hypothesis set at the end of each iteration.

B) **[5 marks]**
What is the worst-case mistake bound for the Halving Algorithm given the hypothesis space described above ? Give an informal derivation of your bound.

Question 6 ANSWER

A) Applying the HALVING ALGORITHM with hypotheses of the form $a \leq x \leq b$ to the data. Start by enumerating the initial hypothesis space[1].

$1 \leq x \leq 5$	$2 \leq x \leq 5$	$3 \leq x \leq 5$	$4 \leq x \leq 5$	$5 \leq x \leq 5$
$1 \leq x \leq 4$	$2 \leq x \leq 4$	$3 \leq x \leq 4$	$4 \leq x \leq 4$	
$1 \leq x \leq 3$	$2 \leq x \leq 3$	$3 \leq x \leq 3$		
$1 \leq x \leq 2$	$2 \leq x \leq 2$			
$1 \leq x \leq 1$				

Now we apply the algorithm to each instance in turn. The algorithm works by running the current instance against each hypothesis in the current hypothesis space. If a hypothesis is consistent with the instance it votes 'POSITIVE' otherwise 'NEGATIVE' (here an instance is consistent with the hypothesis if it is in the range defined by the hypothesis). The class with more votes is the prediction. Then the algorithm checks the actual class of the instance to see if it made a mistake, and it eliminates all of hypotheses that misclassified the instance.

Instance	POSITIVE votes	NEGATIVE votes	Prediction	Actual	Mistake ?
1	5	15	NEGATIVE	NEGATIVE	no

Updated hypothesis space:

$2 \leq x \leq 5$	$3 \leq x \leq 5$	$4 \leq x \leq 5$	$5 \leq x \leq 5$
$2 \leq x \leq 4$	$3 \leq x \leq 4$	$4 \leq x \leq 4$	
$2 \leq x \leq 3$	$3 \leq x \leq 3$		
$2 \leq x \leq 2$			

Get next instance:

Instance	POSITIVE votes	NEGATIVE votes	Prediction	Actual	Mistake ?
2	4	6	NEGATIVE	POSITIVE	yes

Updated hypothesis space:

$2 \leq x \leq 5$
$2 \leq x \leq 4$
$2 \leq x \leq 3$
$2 \leq x \leq 2$

[1]Note: all hypotheses of the form $b \leq x \leq a$ where $a < b$, such as $4 \leq x \leq 2$, have been omitted. Why? These are all equivalent in the sense that they are semantically the same, since they exclude all possible instances. We could have included one of these, but it would not change the results much.

Get next instance:

Instance	POSITIVE votes	NEGATIVE votes	Prediction	Actual	Mistake ?
3	3	1	POSITIVE	POSITIVE	no

Updated hypothesis space:

$2 \leq x \leq 5$
$2 \leq x \leq 4$
$2 \leq x \leq 3$

Get next instance:

Instance	POSITIVE votes	NEGATIVE votes	Prediction	Actual	Mistake ?
4	2	1	POSITIVE	POSITIVE	no

Updated hypothesis space:

$2 \leq x \leq 5$
$2 \leq x \leq 4$

Get next instance:

Instance	POSITIVE votes	NEGATIVE votes	Prediction	Actual	Mistake ?
5	1	1	POSITIVE	NEGATIVE	yes

The last vote required a tie-break. On this occasion it led to a mistake. Nevertheless the algorithm has converged to the correct hypothesis.

Final hypothesis space:

$2 \leq x \leq 4$

B) The worst-case mistake bound for the HALVING ALGORITHM is $\lfloor \log_2 |H| \rfloor$ where H is the hypothesis space. Informally, this can be explained as a kind of "binary chop" procedure. For each instance, the algorithm makes a classification based on a majority vote of the hypotheses in the hypothesis space. If the predicted class is the same as the actual class, there is no mistake, otherwise there is. However, in *either* case all hypotheses that predicted incorrectly are eliminated. So on each mistake, at least half of the hypotheses will be eliminated (because of majority voting).

In this case, the algorithm converged with 2 mistakes, less than the worst-case $\lfloor \log_2(15) \rfloor = 3$.

END OF PAPER