Name of Candidate: ....

Student number: ....

Signature: ....

COMP9417 Machine Learning and Data Mining

Final Examination: SAMPLE QUESTIONS

HERE ARE SEVEN 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]

Learning in Logic

A) **[6 marks]**

Consider the following two clauses:

$C = Q(A, x, B) \lor S(y, B)$ and $C_1 = S(w, B) \lor \neg R(z)$

Using inverse resolution, provide at least one solution for C_2. [Show all substitutions].

B **[8 marks]**

Construct the ***Relative Least General Generalisation (RLGG)*** of two observa-tions: `likes(alan,sushi)` and `likes(alan,curry)`, given the background predicates `food(sushi)` and `food(curry)`. Now suppose you are given two more observations: `likes(bettina,sushi)` and `likes(bettina,curry)`. Will the RLGG of the four obser-vations, given the same background predicates, change ? If you think the answer is yes, give the new RLGG, otherwise give an argument why it will not have changed. [Show all working].

C) **[6 marks]**

Explain how the *generality* order on hypotheses can be expressed for hypotheses that are atoms in first-order logic. Suggest refinement operator for such atoms that could be used to search the hypothesis space.

Question 3 [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 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]

Evaluation of Learning

A) **[6 marks]**
The AUC (area under the ROC curve) measure originated in signal detection theory. For the evaluation of classifier learning on a two-class prediction problem, can you think of a *probabilistic* interpretation of this measure? In this setting, under what conditions does this measure achieve its maximum and minimum values?

B) **[8 marks]**
Suppose we specify the outcome of learning a two-class classifier with the following contingency table:

Actual Class	Predicted Class	
	Yes	No
Yes	TP	FN
No	FP	TN

Two widely-used measures are *true positive rate* or *sensitivity* which is $TPR = \frac{TP}{TP+FN}$, and *true negative rate* or *specificity* which is $TNR = \frac{TN}{TN+FP}$. Explain how accuracy can be calculated as a *weighted average* of TPR and TNR.

C) **[6 marks]**
Once again, for a two-class classification problem, suppose you have the following setting:

  – a data set D has a uniform class distribution, i.e., the class ratio is 1;

  – on a *coverage plot*, two classifiers are evaluated on D and their classification performance is represented by two points C_1 and C_2 on the coverage plot;

  – you observe that C_1 and C_2 can be connected on the coverage plot by a straight line of slope 1.

Which of the classifiers, C_1 or C_2, has greater accuracy? Explain your answer.

Question 6 [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 {true, false, 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 7 [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 G and S sets and those 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.

END OF PAPER