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 classification rules from the tree 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) **[2 marks]**
Explain the difference between the *maximum a posteriori* hypothesis $h_{\text{MAP}}$ and the *maximum likelihood* hypothesis $h_{\text{ML}}$.

B) **[4 marks]**
Would the Naive Bayes classifier be described as a generative or as a discriminative probabilistic model ? Explain your reasoning informally in terms of the conditional probabilities used in the Naive Bayes classifier.

C) **[10 marks]**
Using the multivariate Bernoulli distribution to model the probability of some type of weather occurring or not on a given day, from the following data calculate the probabilities required for a Naive Bayes classifier to be able to decide whether to play or not.

Use pseudo-counts (Laplace correction) to smooth the probability estimates.

| Day | Cloudy | Windy | Play tennis |
|---|---|---|---|
| 1 | 1 | 1 | no |
| 2 | 0 | 1 | no |
| 3 | 1 | 1 | no |
| 4 | 0 | 1 | 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 days ?

| Day | 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 $o$ 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 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.

END OF PAPER