NATIONAL UNIVERSITY OF SINGAPORE

CS3244 - MACHINE LEARNING

(Semester 2: AY2017/18)

Time Allowed: 2 Hours

INSTRUCTIONS TO CANDIDATES

1. This assessment paper contains **FIVE (5)** parts and comprises **THIRTEEN (13)** printed pages, including this page.
2. Answer **ALL** questions as indicated.
3. This is a **OPEN BOOK** assessment.
4. You are allowed to use **NUS APPROVED CALCULATORS**.
5. Please write your **Student Number** below. Do not write your name.

STUDENT NUMBER: ____________________

EXAMINER'S USE ONLY		
Part	Mark	Score
I	18	
II	12	
III	20	
IV	10	
V	20	
TOTAL	80	

In Part I, II, III, IV, and V, you will find a series of structured questions. For each structured question, give your answer in the reserved space in the script.

Part I
Concept Learning 1

(18 points) Structured questions. Answer in the space provided on the script.

1. (8 points) Prove that if an input instance x satisfies none of the members of the general boundary G of the version space $VS_{H,D}$, then x satisfies none of the hypotheses in $VS_{H,D}$.

   Hints: This question is about proving Proposition 4 (in part) on page 32 of the "Concept Learning" lecture slides. You may wish to use the version space representation theorem (VSRT) on page 20 of the "Concept Learning" lecture slides and reproduced below:

$$VS_{H,D} = \{h \in H \mid \exists s \in S\ \exists g \in G\ \ g \geq_g h \geq_g s\}$$

   where S is the specific boundary of $VS_{H,D}$.

Solution:

2. (10 points) Suppose that the target concept c is in the hypothesis space H (i.e., $c \in H$). Give a proof by induction on n that the input instance of every *positive* training example satisfies the hypothesis h_n produced by the FIND-S algorithm after training with a set $D = \{\langle x_k, c(x_k)\rangle\}_{k=1,\ldots,n}$ of n *noise-free* training examples. Note that this question is about proving Proposition 2 (in part) on page 15 of the "Concept Learning" lecture slides. So, you are NOT allowed to use Proposition 2 for the proof in this question.

Solution:

Solution:

Part II
Concept Learning 2

(12 points) Structured questions. Answer in the space provided on the script.

1. (12 points) Consider the hypothetical task of learning the target concept $MLGrade$ to understand the factors affecting the grades of students enrolled in an ML class and the hypothesis space H that is represented by a conjunction of constraints on input attributes, as previously described on page 7 of the "Concept Learning" lecture slides. Each constraint on an input attribute can be a specific value, don't care (denoted by '?'), and no value allowed (denoted by '$\emptyset$'), as previously described on page 5 of the "Concept Learning" lecture slides. Each input instance is represented by the following input attributes:

   - $AttendClass$ (with possible values $Always$, $Sometimes$, $Rarely$),
   - $FinalsGrade$ (with possible values $Good$, $Average$, $Poor$),
   - $ProjectGrade$ (with possible values $Good$, $Average$, $Poor$), and
   - $LoveML$ (with possible values Yes, No).

   For example, a typical hypothesis in H is

   $$\langle ?, Average, ?, Yes \rangle \ .$$

   Trace the CANDIDATE-ELIMINATION algorithm (reproduced below in Fig. 1) for the hypothesis space H given the sequence of positive ($MLGrade = Pass$) and negative ($MLGrade = Fail$) training examples from Table 1 below (i.e., show the sequence of S and G boundary sets).

1. $G \leftarrow$ maximally general hypotheses in H
2. $S \leftarrow$ maximally specific hypotheses in H
3. For each training example d
   - If d is a positive example
     - Remove from G any hypothesis inconsistent with d
     - For each $s \in S$ not consistent with d
       - Remove s from S
       - Add to S all minimal generalizations h of s s.t.
         h is consistent with d, and
         some member of G is more general than h
       - Remove from S any hypothesis that is more general than another hypothesis in S
   - If d is a negative example
     - Remove from S any hypothesis inconsistent with d
     - For each $g \in G$ not consistent with d
       - Remove g from G
       - Add to G all minimal specializations h of g s.t.
         h is consistent with d, and
         some member of S is more specific than h
       - Remove from G any hypothesis that is more specific than another hypothesis in G

Figure 1: CANDIDATE-ELIMINATION algorithm.

Example Student	Input Instances				Target Concept
	$AttendClass$	$FinalsGrade$	$ProjectGrade$	$LoveML$	$MLGrade$
1. $Ryutaro$	$Sometimes$	$Good$	$Poor$	Yes	$Pass$
2. $Haibin$	$Sometimes$	$Good$	$Average$	Yes	$Pass$
3. $Jinho$	$Rarely$	$Average$	$Average$	No	$Fail$
4. $Jingfeng$	$Sometimes$	$Poor$	$Average$	No	$Fail$

Table 1: Positive ($MLGrade = Pass$) and negative ($MLGrade = Fail$) training examples for target concept $MLGrade$.

Solution:

$$
\begin{aligned}
G_0 &= \{\langle ?, ?, ?, ? \rangle\} \\
S_0 &= \{\langle \emptyset, \emptyset, \emptyset, \emptyset \rangle\}
\end{aligned}
$$

$G_1 =$

$S_1 =$

$G_2 =$

$S_2 =$

$S_3 =$

$G_3 =$

$S_4 =$

$G_4 =$

Suppose that the target concept c is in the hypothesis space H (i.e., $c \in H$) and an active learner has already observed the set D of 4 training examples in Table 1 above. State **every** possible input instance (i.e., assuming such a student exists) that the active learner can query next for the 5-th training example to reduce the version space $VS_{H,D}$ by at least half. Note that the active learner does not know the output label $c(x)$ of any input instance x that it has not yet observed.

Hint: Draw the version space $VS_{H,D}$.

Solution:

Part III
Bayesian Inference

(20 points) Structured questions. Answer in the space provided on the script.

You have just come back from dinner at a nice restaurant with your friends and you are feeling a bit queasy. You wonder if you may have food poisoning. Since you and your friends all shared the same dishes, you decide to make some phone calls to see how your friends are feeling.

Before that, you perform a little search on the internet and discover the following probability table relating how a person reports feeling depending on whether or not the food he/she just ate was contaminated:

Person reports feeling	Food is contaminated	Food is not contaminated
Fine	0.1	0.8
Queasy	0.4	0.1
Sick	0.5	0.1

You also discover that the probability that the food in any given restaurant is contaminated is 0.05. You may assume conditional independence between how different persons report feeling given the state of food being contaminated (or not).

Let C and $\neg C$ denote the event of the food being contaminated or not contaminated, respectively. Let F, Q, and S denote the event of the person feeling fine, queasy, or sick, respectively. Finally, let Q_{you} and S_{friend} denote events such as "you feeling queasy" and "friend feeling sick", respectively.

Treat each of the question below **independently**. For example, in question 3 below, assume you have spoken to only that particular friend.

1. (3 points) What is the probability that the food was contaminated if you are feeling queasy? Give your answer up to 3 decimal places. Show the steps of your derivation. **No marks will be awarded for not doing so.**

Solution:

2. (5 points) What is the probability that the food was contaminated if you are feeling queasy and you call one of your friends and find that he is feeling sick? Give your answer up to 3 decimal places. Show the steps of your derivation. **No marks will be awarded for not doing so.**

Solution:

3. (5 points) What is the probability that the food was not contaminated if you are feeling queasy and you call one of your friends and find that she is feeling fine? Give your answer up to 3 decimal places. Show the steps of your derivation. **No marks will be awarded for not doing so.**

Solution:

4. (7 points) Given that you are feeling queasy, how many of your friends would have to be feeling queasy before there is at least a 99% chance that the food was contaminated? Remember to exclude yourself from the final answer. Show the steps of your derivation. **No marks will be awarded for not doing so.**

Solution:

Part IV
Computational Learning Theory

(10 points) Structured questions. Answer in the space provided on the script.

Consider the following setting: Let the sets of all possible input instances, hypotheses, and target concepts/functions be denoted by X, H, and C, respectively. The learner observes a set D of noise-free training examples of the form $\langle x, c(x)\rangle$ of some target concept $c \in C$, where each training instance $x \in X$ is randomly sampled from a fixed probability distribution Q (unknown to the learner) over X to query the teacher for $c(x)$. The learner has to output a hypothesis $h \in H$ to approximate c. We assume c is in the learner's hypothesis space H.

1. (5 points) Derive the value of the true error $error_Q(h)$ of hypothesis h with respect to target concept c and distribution Q if $Q(x) = 0$ for all $x \in X$ such that $h(x) \neq c(x)$. Show the steps of your derivation. **No marks will be awarded for not doing so.**

Solution:

2. (5 points) Let $X' = \{x \in X | h(x) \neq c(x)\}$. Derive the value of the true error $error_Q(h)$ of hypothesis h with respect to target concept c and distribution Q if $Q(x) = 1/(2|X'|)$ for all $x \in X$ such that $h(x) \neq c(x)$. Show the steps of your derivation. **No marks will be awarded for not doing so.**

Solution:

Part V
Neural Networks

(20 points) Structured questions. Answer in the space provided on the script.

1. (4 points) Supposing the weights w_1 and w_2 of a perceptron (see page 6 of "Neural Networks" lecture slides) are both set to the value of **–1**, derive the largest possible range of the values of w_0 that can be set for the perceptron to represent the NAND gate (i.e., NAND(x_1, x_2)). Assume that the inputs x_1 and x_2 and output $o(x_1, x_2)$ of the perceptron are Boolean with the values of **1 or –1**. Show the steps of your derivation. **No marks will be awarded for not doing so.**

Solution:

2. (8 points) Supposing the weights $w_1, w_2, \ldots, w_n$ of a perceptron (see page 6 of "Neural Networks" lecture slides) are all set to the value of **1**, derive the largest possible range of the values of w_0 (in terms of n) that can be set for the perceptron to represent the OR function. That is, the perceptron outputs false if all n Boolean inputs to the perceptron are false, and true otherwise. Assume that the inputs $x_1, x_2, \ldots, x_n$ and output $o(x_1, x_2, \ldots, x_n)$ of the perceptron are Boolean with the values of **1 (i.e., true) or –1 (i.e., false)**. Show the steps of your derivation. **No marks will be awarded for not doing so.**

Solution:

Solution:

3. (8 points) Construct and draw a network of perceptron units with **only one hidden layer (of four units)** that implements (x_1 XOR x_2) XOR x_3 based on the following rules:
   - There should be only one (Boolean) output unit and an input unit for every (Boolean) input.
   - A Boolean is **–1** if false, and **1** if true.
   - The activation function of every (non-input) unit is a -1 to 1 step function (refer to page 6 of the "Neural Networks" lecture slides), including that of the output unit.
   - **Your weights must take on one of the following values: –1, 0, 1, 3.**
   - You don't have to draw edges with weight 0.

   Hint: Observe the truth table of (x_1 XOR x_2) XOR x_3.

Solution:

END OF PAPER