NATIONAL UNIVERSITY OF SINGAPORE

CS3244 - MACHINE LEARNING

(Semester 2: AY2018/19)

Time Allowed: 2 Hours

INSTRUCTIONS TO CANDIDATES

1. This assessment paper contains **SIX (6)** parts and comprises **FOURTEEN (14)** 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	10	
II	10	
III	16	
IV	12	
V	12	
VI	20	
TOTAL	80	

In Part I, II, III, IV, V and VI, 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

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

In a CS3244 lecture, Bryan was huffing and puffing over page 26 of the "Concept Learning" lecture slides. He was trying to explain why, in the CANDIDATE-ELIMINATION algorithm (reproduced below in Fig. 1), if d is a **negative** training example, then each **minimal specialization** h of $g \in G$ (where g is not consistent with d) is not just consistent with d but also consistent with all positive and negative training examples observed thus far. Note from the CANDIDATE-ELIMINATION algorithm in Fig. 1 that some member of S is more specific than h. Can you recall what he was babbling about at that time?

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.

For **both questions below**, you may assume that all hypotheses in G and S are consistent with all positive and negative training examples observed thus far, not including negative training example d.

1. (5 points) Prove formally that each **minimal specialization** h of $g \in G$ (where g is not consistent with d) is consistent with all positive training examples observed thus far. Note from the CANDIDATE-ELIMINATION algorithm in Fig. 1 that h is consistent with d, and some member of S is more specific than h.

Solution:

Solution:

2. (5 points) Prove formally that each **minimal specialization** h of $g \in G$ (where g is not consistent with d) is consistent with all negative training examples observed thus far. Note from the CANDIDATE-ELIMINATION algorithm in Fig. 1 that h is consistent with d, and some member of S is more specific than h.

Solution:

Part II
Concept Learning 2

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

1. (5 points) Let G be the general boundary of the version space $VS_{H,D}$ and $h' \in H$. Give a proof by contradiction that
$$(\forall g \in G \quad g \not\geq_g h') \rightarrow (\forall h \in VS_{H,D} \quad h \not\geq_g h') \, .$$
Hint: You may assume that the transitive property of the $\geq_g$ relation holds. Use the Version Space Representation Theorem (page 20 of the "Concept Learning" lecture slides):
$$VS_{H,D} = \{h \in H \mid \exists s \in S \, \exists g \in G \quad g \geq_g h \geq_g s\} \, .$$

Solution:

2. (5 points) Let S be the specific boundary of the version space $VS_{H,D}$ and $h' \in H$. Give a proof by contradiction that
$$(\forall s \in S \quad h' \not\geq_g s) \rightarrow (\forall h \in VS_{H,D} \quad h' \not\geq_g h) \, .$$
Hint: You may assume that the transitive property of the $\geq_g$ relation holds. Use the Version Space Representation Theorem (page 20 of the "Concept Learning" lecture slides):
$$VS_{H,D} = \{h \in H \mid \exists s \in S \, \exists g \in G \quad g \geq_g h \geq_g s\} \, .$$

Solution:

Part III
Concept Learning 3

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

1. (16 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 $Always$, $Sometimes$, $Rarely$),
   - $MidtermGrade$ (with possible values $Good$, $Average$, $Poor$),
   - $ProjectGrade$ (with possible values $Good$, $Average$, $Poor$), and
   - $LoveML$ (with possible values Yes, No).

   The **difference** $h \setminus h'$ of the hypotheses h and h' is defined as $h \setminus h'(x) = ((h(x) = 1) \wedge (h'(x) = 0))$ for all $x \in X$ and therefore represents the set difference of the sets of input instances represented by h and h'. The **symmetric difference** $h \triangle h'$ of the hypotheses h and h' is defined as $h \triangle h'(x) = (h \setminus h'(x)) \vee (h' \setminus h(x))$ for all $x \in X$ and therefore represents the symmetric difference of the sets of input instances represented by h and h'. Let us define a new hypothesis space H' that consists of all **symmetric differences** of the hypotheses in H. For example, a typical hypothesis in H' is $\langle ?, ?, Good, ? \rangle \triangle \langle ?, Average, ?, Yes \rangle$.

   Trace the CANDIDATE-ELIMINATION algorithm (reproduced below in Fig. 2) 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). You only need to show the **semantically distinct** hypotheses in each boundary set; the hypotheses h and h' are **semantically distinct** iff there exists some $x \in X$ satisfying $h \setminus h'$ or $h' \setminus h$, that is, $\exists x \in X \;\; (h \setminus h'(x) = 1) \vee (h' \setminus h(x) = 1)$.

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 2: CANDIDATE-ELIMINATION algorithm.

Example	Input Instances				Target Concept
Student	$AttendClass$	$MidtermGrade$	$ProjectGrade$	$LoveML$	$MLGrade$
1. $Haibin$	$Sometimes$	$Good$	$Average$	Yes	$Pass$
2. $Yizhou$	$Always$	$Good$	$Average$	Yes	$Pass$
3. $Ryutaro$	$Rarely$	$Good$	$Good$	Yes	$Pass$
4. $TengTong$	$Always$	$Good$	$Good$	Yes	$Pass$
5. $QuocPhong$	$Rarely$	$Poor$	$Good$	Yes	$Fail$

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

Solution:

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

$G_1 =$

$S_1 =$

$G_2 =$

$S_2 =$

$G_3 =$

$S_3 =$

$G_4 =$

$S_4 =$

$S_5 =$

$G_5 =$

Part IV
Expectation Maximization for Estimating 2 Means

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

1. (12 points) Given three instances $x_1 = 4$, $x_2 = 5$, and $x_3 = 6$ from X generated by a mixture of two Gaussian distributions with the same known variance $\sigma^2 = 0.5$, run the Expectation Maximization algorithm for **3 iterations** to estimate the values of the unknown means μ_1 and μ_2 of the two Gaussian distributions. Initialize the values of μ_1 and μ_2 to 4 and 6, respectively. Show the steps of your derivation. **No marks will be awarded for not doing so.** Give your answer up to 6 decimal places.

Solution:

Solution:

Solution:

Part V
Computational Learning Theory

(12 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. (6 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 where $Q(x) = 1/|X|$ for all $x \in X$. Show the steps of your derivation. **No marks will be awarded for not doing so.**

Solution:

2. (6 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 where $Q(x) = 1/|X'|$ if $x \in X'$, and $Q(x) = 0$ otherwise. Show the steps of your derivation. **No marks will be awarded for not doing so.**

Solution:

Part VI
Neural Networks

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

1. (6 points) Suppose that the weights w_0 and w_2 of a perceptron (see page 6 of "Neural Networks" lecture slides) are set to the values of **–1** and **1**, respectively. Derive the **largest** possible range of the values of w_1 that can be set for the perceptron to represent the AND gate (i.e., AND(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. (6 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 NOR function. That is, the perceptron outputs true 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 or –1**. Show the steps of your derivation. **No marks will be awarded for not doing so.**

Solution:

Solution:

3. (8 points) Consider the network of perceptron units in Fig. 3 with a hidden layer of one unit h based on the following **constraints**:

   - There should be only one (Boolean) output unit k and two input units (i.e., one input unit for each of the two (Boolean) input attributes x_1 and x_2).
   - 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.
   - Besides connecting the two input units to the hidden unit h via weights w_1 and w_2, these two input units are also connected to the output unit k via weights w_4 and w_5, as shown in Fig. 3. That is, the two (Boolean) input attributes x_1 and x_2 are also inputs to the output unit k.
   - **The weights w_1, w_2, w_3, w_4, and w_5 must take on one of the following values: –2, –1, 1, 2.**
   - There is **no bias weight** for any unit. Note that the hidden unit h is not a bias input.

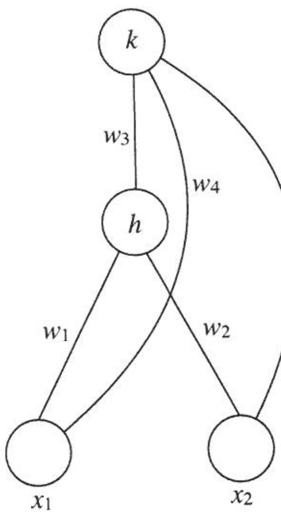

Figure 3: Network of perceptron units.

Yizhou has constructed a dataset of 2 Boolean input attributes x_1 and x_2, and a Boolean output/target concept t_k with the following 4 training examples of the form $d = \langle (x_1, x_2), t_k \rangle$:

$$D = \{d_1 = \langle(-1,-1),1\rangle, d_2 = \langle(-1,1),-1\rangle, d_3 = \langle(1,-1),-1\rangle, d_4 = \langle(1,1),1\rangle\} .$$

Give a hypothesis (i.e., vector of weight values) $(w_1, w_2, w_3, w_4, w_5)$ for the network of perceptron units in Fig. 3 with the above-specified constraints such that your given hypothesis is consistent with D.

Solution:

END OF PAPER