National University of Singapore
School of Computing
CS3244 Machine Learning

Tutorial 1: Concept Learning

Issue: January 17, 2022 Due: January 24, 2022

Important Instructions:

- *Your solutions for this tutorial must be TYPE-WRITTEN.*
- *Make TWO copies of your solutions: one for you and one to be SUBMITTED TO THE TUTOR IN CLASS. Your submission in your respective tutorial class will be used to indicate your CLASS ATTENDANCE. Late submission will NOT be entertained.*
- *Indicate your NAME, STUDENT NUMBER, and TUTORIAL GROUP in your submitted solution.*
- *YOUR SOLUTIONS TO QUESTIONS* **BL 1***,* **TM2.1** *will be GRADED for this tutorial.*
- *You may discuss the content of the questions with your classmates. But everyone should work out and write up ALL the solutions by yourself.*

BL 1 Prove Proposition 1 on page 14 of the "Concept Learning" lecture slides. State any definition or result that you have used from the lecture slides.

Solution.

1. h is consistent with D iff $h(x) = c(x)$ for all $\langle x, c(x)\rangle \in D$, by the definition of a consistent hypothesis.
2. h is consistent with D iff $h(x) = 1$ for all $\langle x, 1\rangle \in D$ and $h(x) = 0$ for all $\langle x, 0\rangle \in D$.
3. h is consistent with D iff every $+$ve training instance satisfies h and every $-$ve training instance does not satisfy h.

TM 2.1 As explained on page 10 of the "Concept Learning" lecture slides, the hypothesis space H for the *EnjoySport* learning task has 973 semantically distinct hypotheses. Note that the representation of a hypothesis in H is still a **conjunction of constraints of input attributes** (page 5 of the "Concept Learning" lecture slides).

(a) Calculate the number $|X|$ of possible input instances and the number $|H|$ of semantically distinct hypotheses with the addition of a new input attribute *WaterCurrent* that can take on the values *Light*, *Moderate*, or *Strong*.

(b) More generally, calculate the number of possible input instances and the number of semantically distinct hypotheses with the addition of a new input attribute that takes on k possible values. Your answers should be expressed in terms of k, $|X|$, and $|H|$ where $|X|$ and $|H|$ denote, respectively, the number of possible input instances and the number of semantically distinct hypotheses BEFORE the addition of the new input attribute.

Solution.

(a) $|X| = 3 \times 2 \times 2 \times 2 \times 2 \times 2 \times 3 = 288$. $|H| = 1 + 4 \times 3 \times 3 \times 3 \times 3 \times 3 \times 4 = 3889$.

(b) Adding a new input attribute with k possible values yields $k|X|$ possible instances and $(1 + (k+1)(|H| - 1)) = |H|(k+1) - k$ possible hypotheses.

TM 2.2 (a) Give the sequence of S and G boundary sets computed by the Candidate-Elimination algorithm if it is given the sequence of training examples from the table on page 4 of the "Concept Learning" lecture slides in **reverse** order.
(b) Although the final version space will be the same regardless of the sequence of examples (why?), the sets S and G computed at intermediate stages will, of course, depend on this sequence. Can you come up with ideas of heuristics for ordering the training examples to minimize the sum of the sizes of these intermediate S and G sets for the H used in the $EnjoySport$ task?

Solution.

(a)

$$\begin{aligned}
S_0: &\ \{\langle \emptyset, \emptyset, \emptyset, \emptyset, \emptyset, \emptyset \rangle\} \\
S_1, S_2: &\ \{\langle Sunny, Warm, High, Strong, Cool, Change \rangle\} \\
S_3: &\ \{\langle Sunny, Warm, High, Strong, ?, ? \rangle\} \\
S_4: &\ \{\langle Sunny, Warm, ?, Strong, ?, ? \rangle\} \\
\\
G_3, G_4: &\ \{\langle Sunny, ?, ?, ?, ?, ? \rangle, \langle ?, Warm, ?, ?, ?, ? \rangle\} \\
G_2: &\ \{\langle Sunny, ?, ?, ?, ?, ? \rangle, \langle ?, Warm, ?, ?, ?, ? \rangle, \langle ?, ?, ?, ?, Cool, ? \rangle\} \\
G_0, G_1: &\ \{\langle ?, ?, ?, ?, ?, ? \rangle\}
\end{aligned}$$

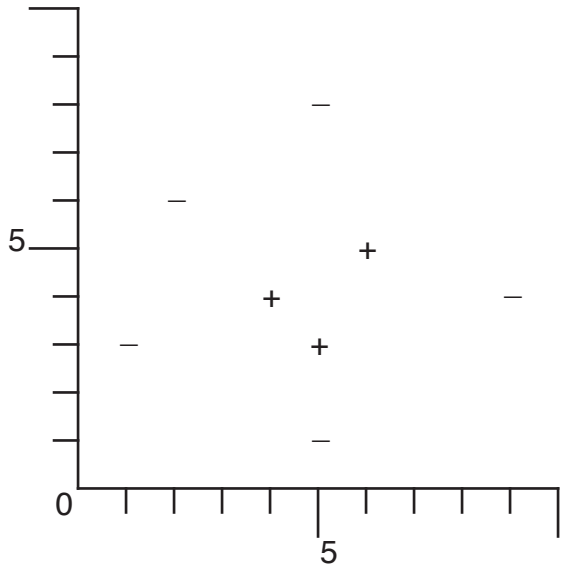
(b) To answer why the sequence/order does not matter, the algorithm is designed to find **all** maximally general and specific hypotheses, hence the final version space. Also, the definitions of G and S (respectively, sets of maximally general and specific hypotheses) are **independent** of the sequence/order of the training examples in the dataset.

Heuristic 1. Select all the $+$ve training examples first, followed by the $-$ve examples. Using the hypothesis representation of conjunction of attribute constraints in H, the set S never

grows in size since it maintains only one maximally specific hypothesis in each iteration. After training with all the +ve examples, the resulting maximally specific hypothesis in S can reduce the number of specializations of g added to G during training with −ve examples. Thus, this ordering will help to minimize the intermediate set sizes.

TM 2.4 Consider the instance space consisting of integer points in the x, y plane and the set of hypotheses H consisting of rectangles. More precisely, hypotheses are of the form $a \leq x \leq b$, $c \leq y \leq d$, where a, b, c, and d can be any integers.

(a) Consider the version space with respect to the set of positive (+) and negative (−) training examples shown below. What is the S boundary of the version space in this case? Write out the hypotheses and draw them in on the diagram.



(b) What is the G boundary of this version space? Write out the hypotheses and draw them in.

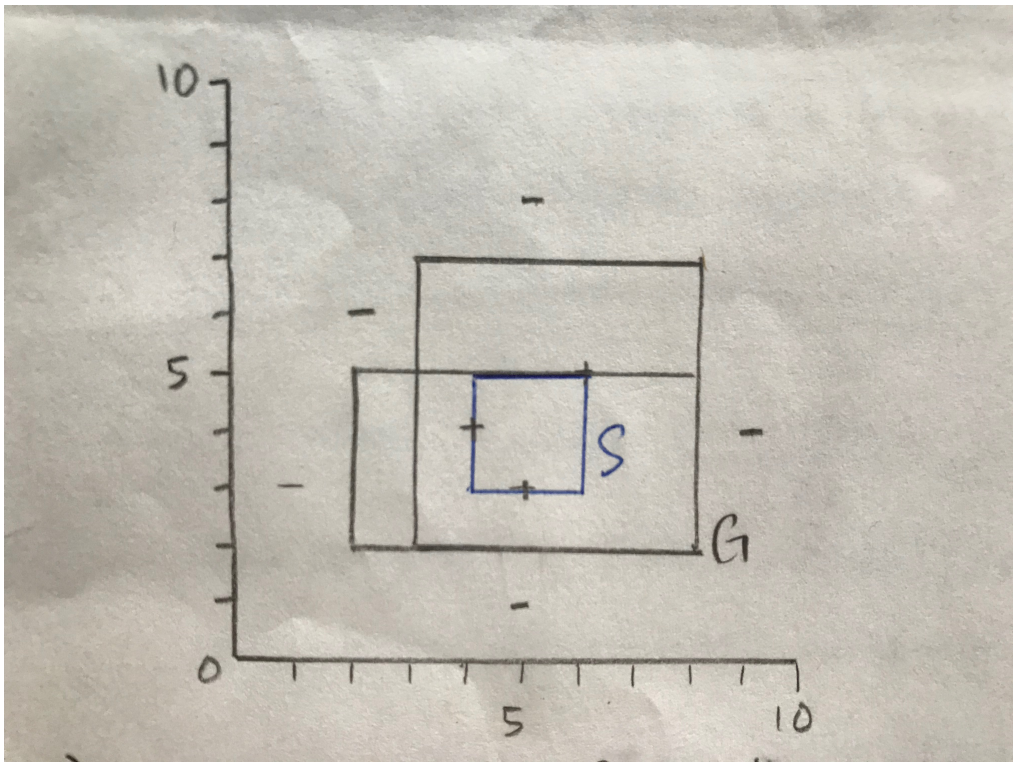
(c) Suppose that the learner may now suggest a new x, y instance and ask the trainer for its classification. Suggest a query guaranteed to reduce the size of the version space, regardless of how the trainer classifies it. Suggest one that will not.

(d) Now assume that you are a teacher attempting to teach a particular target concept (e.g., $3 \leq x \leq 5$, $2 \leq y \leq 9$). What is the smallest number of training examples you can provide so that the Candidate-Elimination algorithm will perfectly learn the target concept?

Solution.

(a) The S boundary is $\{(4 \leq x \leq 6, 3 \leq y \leq 5)\}$.

(b) The G boundary is $\{(2 \leq x \leq 8, 2 \leq y \leq 5), (3 \leq x \leq 8, 2 \leq y \leq 7)\}$.



(c) Any query that comes from inside G but outside S is guaranteed to reduce the version space.

Any query that comes from inside S or from outside G will never reduce the version space.

(d) Perfectly learning the target concept means that the S and G boundaries must both be equal to the actual target concept at the end of learning. Thus, the problem has two parts: What is the smallest number of examples to make S the target concept and what is the smallest number of examples to make G the target concept?

The smallest number of examples necessary to make S the target concept is 2, and these must be the opposite corners of the rectangle.

The smallest number of examples necessary to make G the target concept is 4, and these $-$ve examples must lie on four sides of the target rectangle.

Thus, the total number of examples needed is 6.