16s1: COMP9417 Machine Learning and Data Mining

Lecture: Concept Learning (1)
Exercises: Some Version Spaces exercises from Mitchell's book [1].
Version: with answers
Status: Revised
Last revision: Mon Mar 14 16:41:41 AEDT 2016

Introduction

Try and answer these questions on concept learning taken from Chapter 2 of the book by Tom Mitchell [1].

Exercise 2.1

Question Explain why the size of the hypothesis space in the *EnjoySport* learning task is 973. How would the number of possible instances and possible hypotheses increase with the addition of the attribute *Watercurrent*, which can take on the values *Light*, *Moderate* or *Strong* ? More generally, how does the number of possible instances and hypotheses grow with the addition of a new attribute A that takes on k possible values?
Answer Instance space size will be size of original instance space times 3, i.e., $96 \times 3 = 288$. Size of hypothesis space will increase to $973 \times 4 = 3892$. More generally, the sizes of instance and hypothesis spaces are as follows.

Let $v(a)$ be the set of values for attribute a and $|s|$ be the size of set s. For m attributes $a_i, \ldots, a_m$ the size of the instance space is:

$$\prod_{i=1}^{m} |v(a_i)|$$

For the hypothesis language used in the *EnjoySport* example the size of the hypothesis space is:

$$1 + \prod_{i=1}^{m} (|v(a_i)| + 1)$$

Exercise 2.2

Question A trace of the Candidate-Elimination algorithm on Slides 24–25 of the lecture "Concept Learning (1)" running on the the sequence of training examples below from Slide 7 of the lecture (which is also Table 2.1 on page 21 of [1]) is given on Slides 26–30 of the lecture. First ensure you can apply Candidate-Elimination to reproduce this trace.

Now apply Candidate-Elimination to the same examples but *in reverse order*.

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 for ordering the training examples to minimize the sum of the sizes of these intermediate S and G sets for the hypothesis space H used in the *EnjoySport* example?

Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

Answer

S_0	$\{\langle\emptyset,\emptyset,\emptyset,\emptyset,\emptyset,\emptyset\rangle\}$
G_0	$\{\langle ?,?,?,?,?,?\rangle\}$
Ex. 1	$\langle$*Sunny Warm High Strong Cool Change* $\rangle$, EnjoySport = yes
S_1	$\{\langle Sunny, Warm, High, Strong, Cool, Change\rangle\}$
G_1	$\{\langle ?,?,?,?,?,?\rangle\}$
Ex. 2	$\langle$*Rainy Cold High Strong Warm Change* $\rangle$, EnjoySport = no
S_2	$\{\langle Sunny, Warm, High, Strong, Cool, Change\rangle\}$
G_2	$\{\langle Sunny,?,?,?,?,?\rangle,$
	$\langle ?, Warm,?,?,?,?\rangle,$
	$\langle ?,?,?,?,?, Cool\rangle\}$
Ex. 3	$\langle$*Sunny Warm High Strong Warm Same* $\rangle$, EnjoySport = yes
S_3	$\{\langle Sunny, Warm, High, Strong,?,?\rangle\}$
G_3	$\{\langle Sunny,?,?,?,?,?\rangle,$
	$\langle ?, Warm,?,?,?,?\rangle\}$
Ex. 4	$\langle$*Sunny Warm Normal Strong Warm Same* $\rangle$, EnjoySport = yes
S_4	$\{\langle Sunny, Warm,?, Strong,?,?\rangle\}$
G_4	$\{\langle Sunny,?,?,?,?,?\rangle,$
	$\langle ?, Warm,?,?,?,?\rangle\}$

Final version space is the same despite example ordering since by definition it is the set of consistent hypotheses.
One strategy to minimise the cumulative sizes of the S and G sets through all iterations is to first add all the positive examples, then the negatives. The first stage is like FIND-S and will get to the final S boundary without growing the G set. Then when the negative examples are added, the G set will converge without making any specializations that are inconsistent with all the positive examples (which would unnecessarily grow the G set).

Exercise 2.3

Question Consider again the *EnjoySport* learning task and the hypothesis space H described on slides 8–10 of the lecture. Let us define a new hypothesis space H' that consists of all pairwise disjunctions of the hypotheses in H. For example, a typical hypothesis in H' is

$$(?, \mathit{Cold}, \mathit{High}, ?, ?, ?) \vee (\mathit{Sunny}, ?, \mathit{High}, ?, ?, \mathit{Same})$$

Trace the Candidate-Elimination algorithm for the hypothesis space H' given the sequence of training examples as in the table above (i.e., show the sequence of S and G boundary sets.)

Answer In this exercise the hypothesis space become the set of *pairwise disjunctions* of hypotheses in the hypothesis space used in the previous exercise (2.2) on the *EnjoySport* data.

Call this hypothesis space H_1 and the new hypothesis space H_2. You can start by writing out some of the types of hypotheses h in each of H_1 and H_2 to see some relationships. For example:

H_1 $\{\langle\emptyset,\emptyset,\emptyset,\emptyset,\emptyset,\emptyset\rangle$,
all h with only constant values, i.e., instances,
all h with exactly one variable or “don't care” value, i.e., '?',
all h with exactly two variables,
. . .,
$\langle ?, ?, ?, ?, ?, ?\rangle\}$

For example:

H_2 $\{\langle\emptyset,\emptyset,\emptyset,\emptyset,\emptyset,\emptyset\rangle \vee \langle\emptyset,\emptyset,\emptyset,\emptyset,\emptyset,\emptyset\rangle$
$\langle\emptyset,\emptyset,\emptyset,\emptyset,\emptyset,\emptyset\rangle \vee I$ where I is an instance,
$\langle\emptyset,\emptyset,\emptyset,\emptyset,\emptyset,\emptyset\rangle \vee h$ where h is a 1-variable hypothesis,
$\langle\emptyset,\emptyset,\emptyset,\emptyset,\emptyset,\emptyset\rangle \vee h$ where h is a 2-variable hypothesis,
. . .
all disjuncts $I_1 \vee I_2$ where I_1 and I_2 are instances,
all disjuncts of an instance with a 1-variable h,
all disjuncts of an instance with a 2-variable h,
. . .
all disjuncts of a 1-variable h with a 1-variable h,
all disjuncts of a 1-variable h with a 2-variable h,
. . .
all disjuncts of a 2-variable h with a 2-variable h,
all disjuncts of a 2-variable h with a 3-variable h,
. . .,
$\langle ?, ?, ?, ?, ?, ?\rangle \vee \langle ?, ?, ?, ?, ?, ?\rangle\}$

Some observations:

1. any pair of hypotheses $h_i \vee h_j$ is equivalent to the set union of the hypotheses
2. any pair of hypotheses $h_i \vee h_j$ where h_i is *more general than* h_j is equivalent to h_i

We can see that the Candidate Elimination operations affected by the change in hypothesis space will be in finding the set of “next-most-general” or “next-most-specific” hypotheses for a given hypothesis. There are now additional options for generating a set of new hypotheses from a hypothesis h:
Here are the set of examples.

Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

S_0 $\{\langle\emptyset,\emptyset,\emptyset,\emptyset,\emptyset,\emptyset\rangle \vee \langle\emptyset,\emptyset,\emptyset,\emptyset,\emptyset,\emptyset\rangle\}$
G_0 $\{\langle ?,?,?,?,?,?\rangle \vee \langle ?,?,?,?,?,?\rangle\}$

Ex. 1 $\langle \mathit{Sunny\ Warm\ Normal\ Strong\ Warm\ Same}\ \rangle$, EnjoySport = yes
S_1 $\{\langle \mathit{Sunny}, \mathit{Warm}, \mathit{Normal}, \mathit{Strong}, \mathit{Warm}, \mathit{Same}\rangle\langle\emptyset,\emptyset,\emptyset,\emptyset,\emptyset,\emptyset\rangle\}$
G_1 $\{\langle ?,?,?,?,?,?\rangle \vee \langle ?,?,?,?,?,?\rangle\}$

Ex. 2 $\langle \mathit{Sunny\ Warm\ High\ Strong\ Warm\ Same}\ \rangle$, EnjoySport = yes
S_2 $\{\langle \mathit{Sunny}, \mathit{Warm}, \mathit{Normal}, \mathit{Strong}, \mathit{Warm}, \mathit{Same}\rangle \vee \langle \mathit{Sunny}, \mathit{Warm}, \mathit{High}, \mathit{Strong}, \mathit{Warm}, \mathit{Same}\rangle\}$
G_2 $\{\langle ?,?,?,?,?,?\rangle \vee \langle ?,?,?,?,?,?\rangle\}$

Ex. 3 $\langle \mathit{Rainy\ Cold\ High\ Strong\ Warm\ Change}\ \rangle$, EnjoySport = no
S_3 $\{\langle \mathit{Sunny}, \mathit{Warm}, \mathit{Normal}, \mathit{Strong}, \mathit{Warm}, \mathit{Same}\rangle \vee \langle \mathit{Sunny}, \mathit{Warm}, \mathit{High}, \mathit{Strong}, \mathit{Warm}, \mathit{Same}\rangle\}$
G_3 $\{\langle \mathit{Sunny},?,?,?,?,?\rangle \vee \langle ?,\mathit{Warm},?,?,?,?\rangle,$
$\langle \mathit{Sunny},?,?,?,?,?\rangle \vee \langle ?,?,\mathit{Normal},?,?,?\rangle,$
$\langle \mathit{Sunny},?,?,?,?,?\rangle \vee \langle ?,?,?,?,?,\mathit{Same}\rangle,$
$\langle ?,\mathit{Warm},?,?,?,?\rangle \vee \langle ?,?,\mathit{Normal},?,?,?\rangle,$
$\langle ?,\mathit{Warm},?,?,?,?\rangle \vee \langle ?,?,?,?,?,\mathit{Same}\rangle,$
$\langle ?,?,\mathit{Normal},?,?,?\rangle \vee \langle ?,?,?,?,?,\mathit{Same}\rangle\}$

Ex. 4 $\langle \mathit{Sunny\ Warm\ High\ Strong\ Cool\ Change}\ \rangle$, EnjoySport = yes
S_4 $\{\langle \mathit{Sunny}, \mathit{Warm}, ?, \mathit{Strong}, \mathit{Warm}, \mathit{Same}\rangle \vee \langle \mathit{Sunny}, \mathit{Warm}, \mathit{High}, \mathit{Strong}, \mathit{Cool}, \mathit{Change}\rangle\}$
G_4 $\{\langle \mathit{Sunny},?,?,?,?,?\rangle \vee \langle ?,\mathit{Warm},?,?,?,?\rangle,$
$\langle \mathit{Sunny},?,?,?,?,?\rangle \vee \langle ?,?,\mathit{Normal},?,?,?\rangle,$
$\langle \mathit{Sunny},?,?,?,?,?\rangle \vee \langle ?,?,?,?,?,\mathit{Same}\rangle,$
$\langle ?,\mathit{Warm},?,?,?,?\rangle \vee \langle ?,?,\mathit{Normal},?,?,?\rangle,$
$\langle ?,\mathit{Warm},?,?,?,?\rangle \vee \langle ?,?,?,?,?,\mathit{Same}\rangle\}$

Clearly, there is still a strong language bias, but comparing with the solution for H_1 in the slides we notice that:

- S_4 is now more specific
- G_4 now contains more hypotheses due to the larger hypothesis space

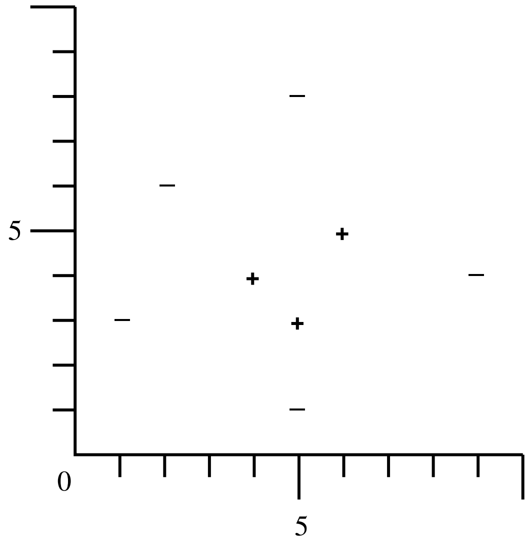

Figure 1: Training set instances.

Exercise 2.4

Question 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 in Figure 1. 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 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 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 ?
Answer
(a) The S set of hypotheses is: $\{(4 \leq x \leq 6, 3 \leq y \leq 5)\}$. This is shown in Figure 2.
(b) The G set of hypotheses is: $\{G1, G2\}$, where $G1 = (2 \leq x \leq 8, 2 \leq y \leq 5)$ and $G2 = (3 \leq x \leq 8, 2 \leq y \leq 7)$. Note that $G1 \not>_g G2$ and vice versa. This is shown in Figure 3. The version space with all G and S set hypotheses is shown in Figure 4.
(c) Suggesting an instance x that is included in a hypothesis in the G set and not in any hypothesis in the S set is guaranteed to reduce the version space (VS) – if x is positive (in the target concept)

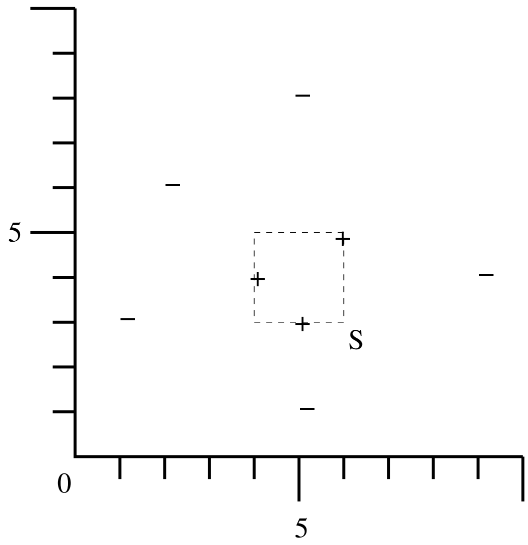

Figure 2: *S* set boundary of version space contains a single hypothesis (shown by dashed line rectangle). Instances are (x,y) points labelled as positive ('+') or negative ('-') examples of the target concept.

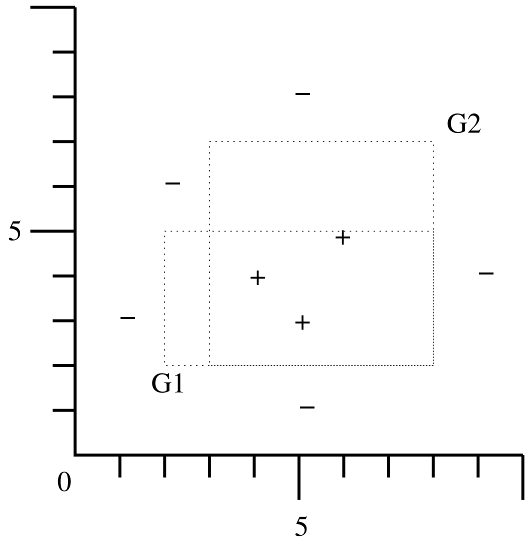

Figure 3: *G* set boundary of version space contains two hypotheses (shown by dotted line rectangles).

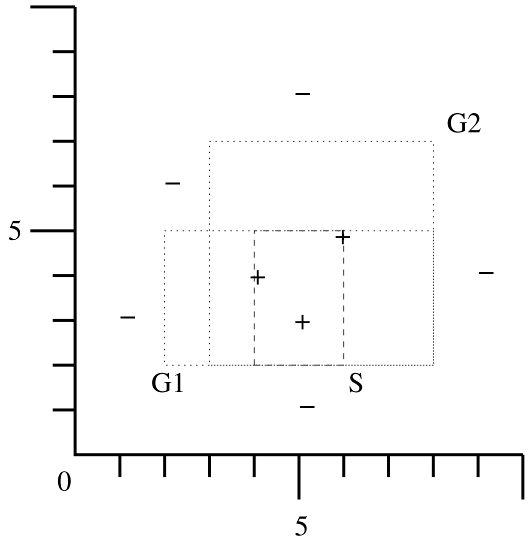

Figure 4: Combined G set (dotted rectangles) and S set (dashed rectangle) hypotheses.

then hypotheses in S will be generalised, reducing the VS, and if x is negative then hypotheses in G will be specialised, also reducing the VS.

Assuming the target hypothesis is in the VS, suggesting an instance x that is included in a hypothesis in the S set and not in any hypothesis in the G will not reduce the VS – whether x is positive or negative, it's classification is already known. The reasoning is similar to that used to show a hypothesis learned by FIND-S does not cover a negative example (Slide 14).

(d) To perfectly learn the target concept the G and S sets must converge to a hypothesis that is equivalent to the target concept. For the rectangle hypothesis language, any target hypothesis will be the specification of a rectangle, e.g., $(3 \leq x \leq 5, 2 \leq y \leq 9)$.

The S set can be specified by giving a set of positive examples to define the minimum and maximum x and y values defining the target rectangle. At least two examples are required, i.e., in this case, either $(3, 2)$ and $(5, 9)$, or $(3, 9)$ and $(5, 2)$.

The G set can be specified by giving a set of negative examples to define the "outside" of the target rectangle. This requires at least four examples, one to define each of the four sides. There are many possible points that could be chosen, but in this case we could have, e.g., $\{(2, 3), (4, 10), (6, 4), (3, 1)\}$.

Therefore the smallest number of training examples for CANDIDATE-ELIMINATION to learn a correct hypothesis in this hypothesis language is six.

References

[1] T. Mitchell. *Machine Learning.* McGraw-Hill, New York, 1997.