CS478 – Homework 1 – Solutions

Alin Dobra

March 6, 2000

Problem 1.2(5pt)

For this problem a treatment similar to the one in Chapter 1 in the text was expected.

- 1 point for the problem
- 1 point for the performance measure
- 1 point for the target function
- 1 point for the target function representation
- 1 point for pointing out the tradeoffs (0.5 for tradeoffs in function representation, most of you failed to mention that a linear representation might be too restrictive).

  If one of the aspects was not clear I took points accordingly.

Problem 2.2

The succession of boundaries is:

$S_1 = < \emptyset, \emptyset, \emptyset, \emptyset, \emptyset, \emptyset >$

$G_1 = <?, ?, ?, ?, ?, ? >$

$S_2 = <$ Sunny, Warm, High, Strong, Cool, Change $>$

$G_2 = G_1$

$S_3 = S_2$

$G_3 = \{<$ Sunny, ?, ?, ?, ?, ? $>, <?,$ Warm, ?, ?, ?, ?, ? $>, <?, ?, ?, ?,$ Cood, ? $>\}$

$S_4 = <$ Sunny, Warm, High, Strong, ?, ? $>$

$G_4 = \{<$ Sunny, ?, ?, ?, ?, ? $>, <?,$ Warm, ?, ?, ?, ?, ? $>\}$

$S_5 = <$ Sunny, Warm, ?, Strong, ?, ? $>$

$G_5 = G_4$

The same kind of reasoning as in the text can be used to justify this solution.

If all the positive examples are considered first the G boundary remains G_1 (the most general boundary) and the S boundary becomes S_5 (the S set contains all the time only one element). If the only negative case is considered afterwards G becomes G_4 (that has 2 elements) without going through G_3 (that has 3 elements), thus the size of the version space is 3 or less all the time.

Grading: 8 points for the first part. I subtracted 1.5 points for each of the 5 different sets in the above sequence (S_2,G_3,S_4, G_4,S_5) and 0.5 for initial situation (S_1, G_1). The last part was worth 2 points (1 for a good ordering, 1 for some justification).

Problem 2.4(15pt)

Hypotheses are rectangles, the $\leq_g$ relation is the inclusion relation on rectangles.

a)

There is only one hypotheses in the S boundary: ($4 \leq x \leq 6$,$3 \leq y \leq 5$). Note that this rectangle is included in all the rectangles that include all the

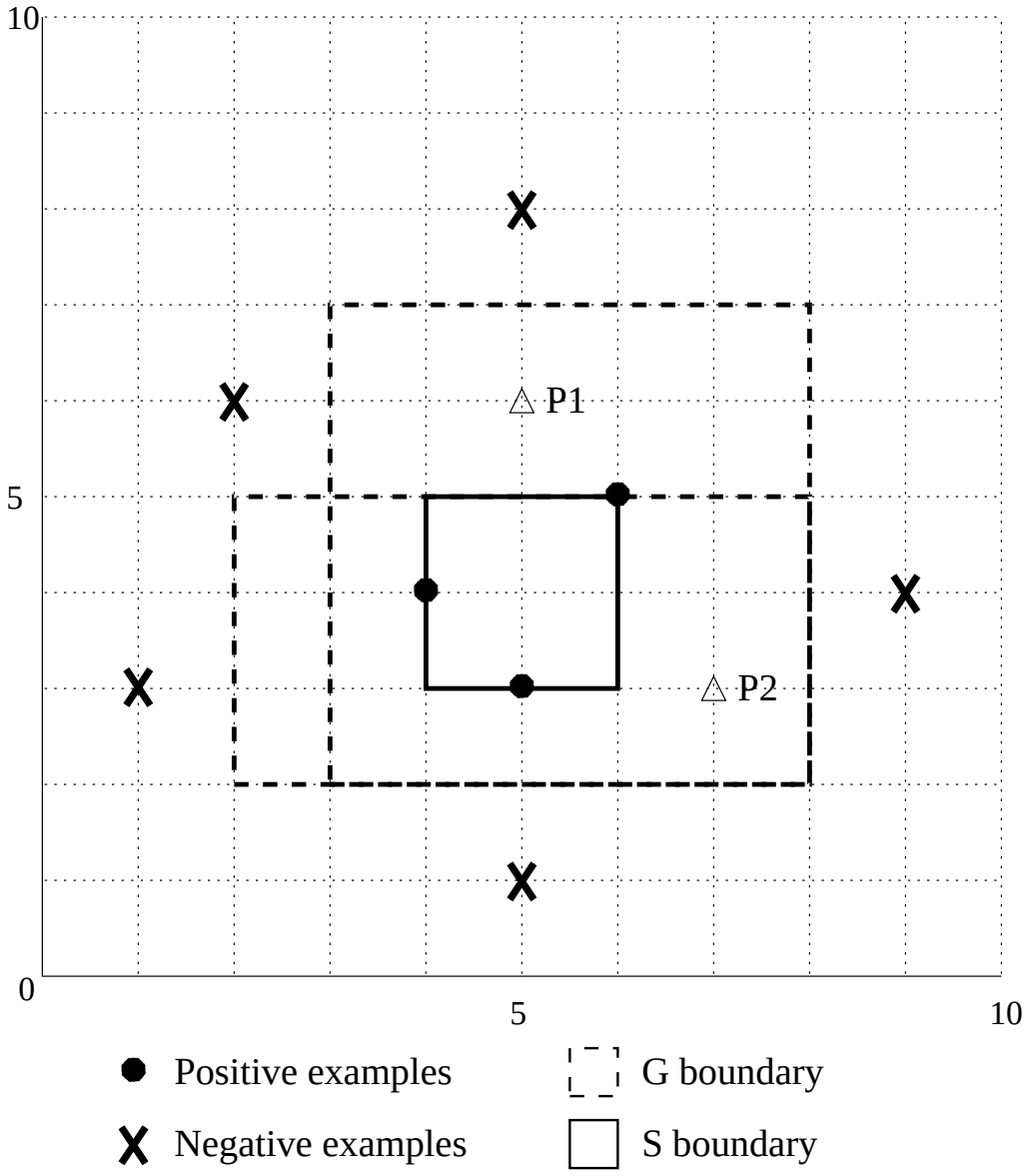

Figure 1: The S and G boundaries for problem 2.4

positive cases (so is the most specific one).

Grading: This part is 3 points. I usually accepted solutions without explanation (is the output of candidate-elimination).

b)

The rectangles in G are as big as possible, contain all the positive cases and don't contain any negative case. Looking at Figure 1 is easy to see that rectangles ($2 \leq x \leq 8$,$2 \leq y \leq 5$) and ($3 \leq x \leq 8$,$2 \leq y \leq 7$) have this property, are not related by $\leq_g$ relation so are the rectangles in G.

Grading: 2 points for each of the two elements in G. If one was missing I took of 2 points

c)

If the query is anywhere between the element in the S bound and any of the two elements (say (5,6)) than if the label is positive the S bound has to grow to incorporate that point, if is negative one or both rectangles in G have to shrink to avoid the point, so the dimension of the version space (the number of hypotheses in the version space) is reduced since at least one modification is made (at least a hypothesis is excluded).

If the query is inside the S bound then it will leave the version space unchanged if is positive, but will reduce the version space to the empty set (so the size goes down to 0) if is negative. Similarly if the query is outside the G bound, if the label is negative anything stays the same, if is positive the date is inconsistent and the version space is empty (0 size). Most of you noticed the first part of the reasoning but failed to observe that the size is shrunk for such a query also (becomes 0).

So there is no point that leaves the version space unchanged no matter what labeling it has.

Grading: 2 points for a point that leaves anything unchanged, 2 point for one that keeps things the same. I took of 1 point in each part if no explanation was provided and 0.5 points if you failed to observed that for a point in S (or outside G) the version space shrinks also.

d)

(4 points) The goal is to have only one hypothesis in both S and G that perfectly describes the concept ($3 \leq x \leq 5$,$2 \leq y \leq 9$). Placing two positive examples at (3,2) and (5,9) ensures that any hypothesis must include the rectangle ($3 \leq x \leq 5$,$2 \leq y \leq 9$) and S contains only this hypothesis.

The hypothesis in G should include $3 \leq x \leq 5$,$2 \leq y \leq 9$. To ensure tha G contains only this rectangle we can place negative examples (elements of G should not include negative examples) to get the desirable G. One way to do this is to place 4 negative points as in the Figure 2, just outside the edges of rectangle $3 \leq x \leq 5$,$2 \leq y \leq 9$. Since any element of G should contain the element in S and none of the negative points, G contains only rectangle $3 \leq x \leq 5$,$2 \leq y \leq 9$ (this rectangle cannot be extended in any way without avoiding a negative point). If one of the points in this configuration is missing, such an expansion is possible so elements of G are not maximal, so they are not elements of G (some other rectangle more general, i.e. including this one, is in G).

So the answer is that you need 6 points (2 positive, 4 negative) to perfectly learn the concept.

Some of you suggested that placing just two negative examples in the outside corners of the rectangle we want to learn solves the problem (triangles in the figure). This is not the case since $G = (0 \leq 5, 2 \leq 10), (3 \leq 10, 0 \leq 9)$ not the required rectangle. I took off 0.5 points for this solution.

An other wrong solution is to place 4 negative examples just outside the corners of the rectangle. This doesn't work either since $G = (0 \leq 10, 2 \leq 9), (3 \leq 5, 0 \leq 10)$. For this case I took 0.5 points off also.

If the solution was non-optimal I took 0.5 points off. If the solution was wrong (but not in the subtle way described before) I took 1 point off.

I gave 2 points for explaining what is happening in some detail (things are far from trivial so an explanation is required).

Some of you told me where to put the points but didn't mention the labeling so I took 1 point off.

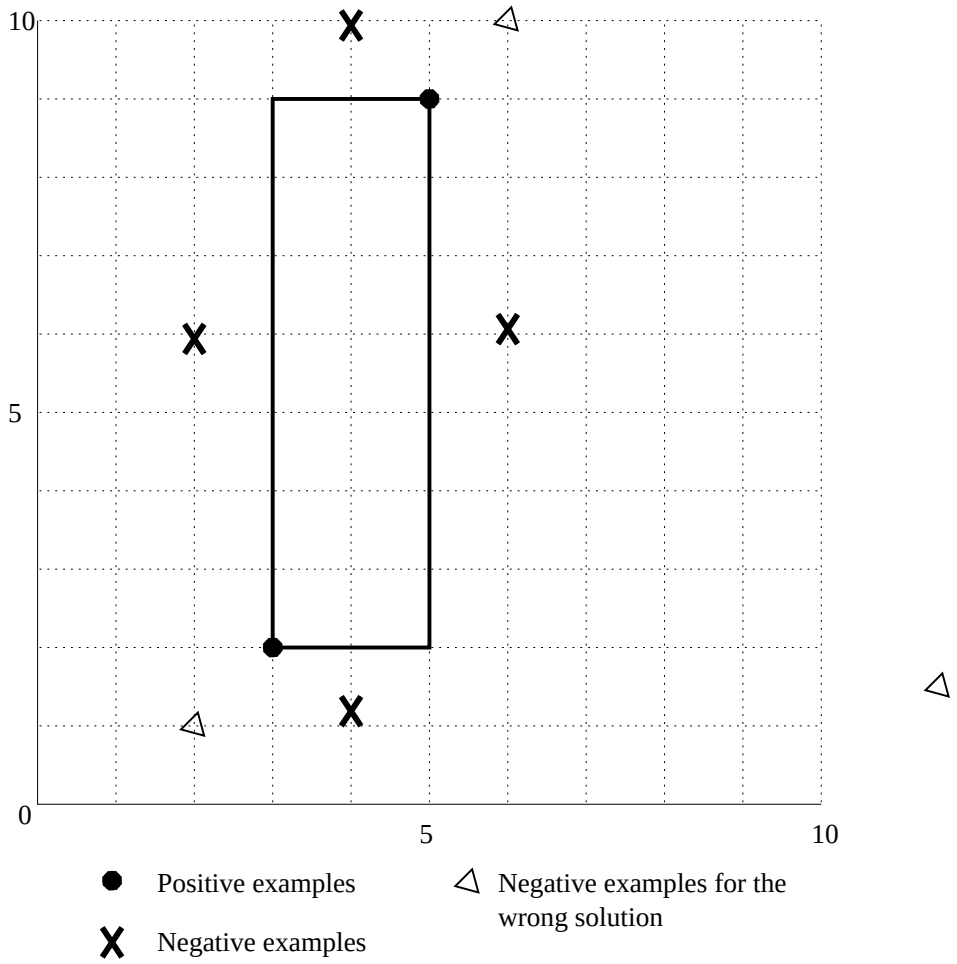

Figure 2: The placement of examples for problem 2.4 d)