This is a possible solution to the homework, although there may be other correct responses to some of the questions. The questions are repeated in this font, while answers are `in a monospaced font.`

1 Version Spaces

Consider the instance space X of points $\vec{x} = (x_1, x_2)$ with integer coordinates $([0..10], [0..10]) \subset N^2$. The task is to learn a concept $c : X \rightarrow \{0, 1\}$ which can be described as a rectangle $((a, b), (c, d)) \in N^2 \times N^2$. (a, b) is the left top corner of the rectangle, (c, d) is the lower right corner. An example is labeled positive, if it lies inside the rectangle or on it's boundary. It is negative, if it lies outside. The hypothesis space H are all rectangles over the instance space X.

Note that in this setting, a hypothesis $h = ((a, b), (c, d))$ can be generalized by decreasing a or b and/or increasing c or d. Similarly it can be made more specific by increasing a or b or decreasing c or d.

Part a (15 POINTS)

Suppose you have the hypothesis $h = ((4, 5), (4, 5))$, after seeing the positive training example $(4, 5)$. If you then see the positive training example $x = (6, 3)$, what is the smallest generalization of h that also accepts x?

`If you use axes with the origin in the usual location, the smallest generalization that covers both examples is` $h = ((4, 5),\ (6, 3))$`.`

`Alternatively, you can have the origin in the top left corner, as implied by the question, which gives the smallest generalization of` $h = ((4, 3),\ (6, 5))$`.`

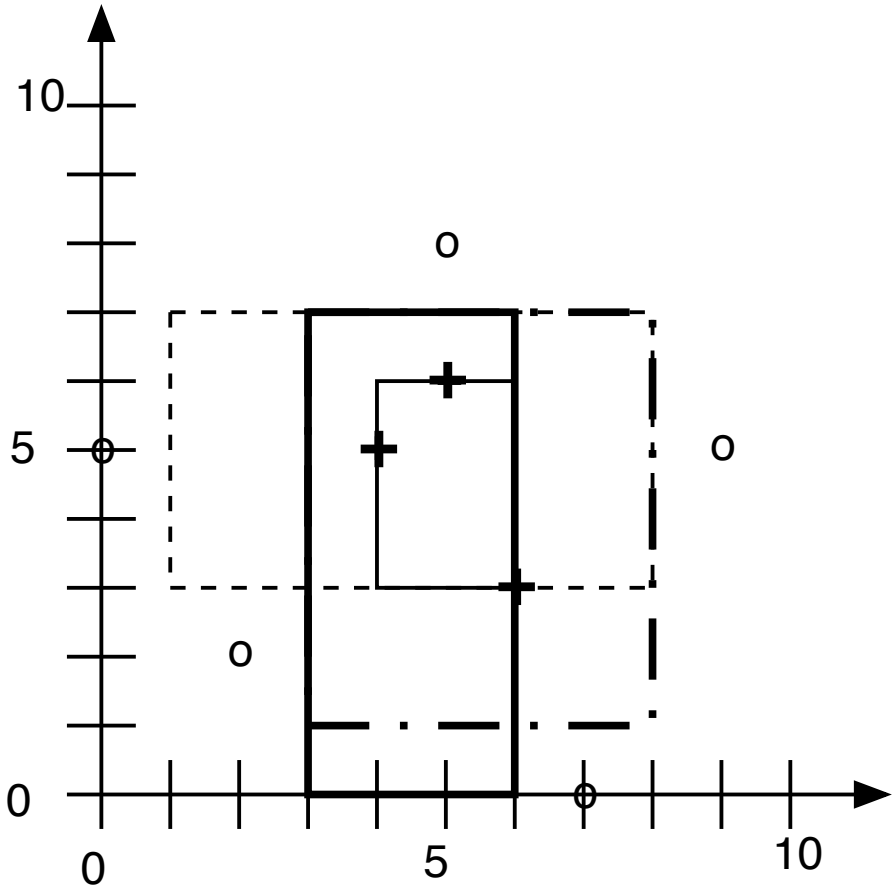
Part b (15 POINTS)

Consider the following training set in $X \times \{0, 1\}$, where the 0 or 1 is the label of each data point (0 indicates negative examples, and 1 indicates positive examples):

$\{((0, 5),\ 0),\ ((4, 5),\ 1),\ ((2, 2),\ 0),\ ((9, 5),\ 0),\ ((6, 3),\ 1),\ ((5, 6),\ 1),\ ((7, 0),\ 0),\ ((5, 8),\ 0)\}$

For this training set, what is the S boundary of the version space? What is the G boundary of the version space? Write out the hypotheses and draw them into a diagram showing the points as well as the boundaries of the version space.

Drawing the datapoints with an origin in the lower left corner, you get the following diagram. Note the S boundary (smallest, thin solid rectangle) and G boundaries (three other rectangles).



There are three G boundaries, that is three rectangles that contain only positive points and cannot be generalized without including any negative points. None of these rectangles is completely contained in another.

We can write the boundaries to be:

$$S = \{((4,6),\ \ (6,3)\}$$
$$G = \{((1,7),\ \ (8,3)),\ \ ((3,7),\ \ (8,1)),\ \ ((3,7),\ \ (6,0))\}$$

Many of the figures were just barely big enough to read. There is no reason to make them too small!

Part c (15 POINTS)

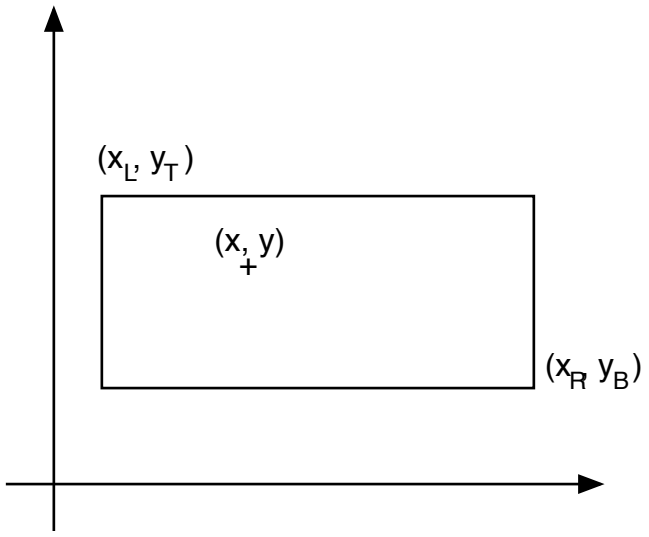
Imagine the learner does not get specific training examples, but instead can propose examples and ask for the correct label. How should the learner pick the example it asks to be labeled

next? Assume that one positive training example is given initially. Describe a strategy that will lead the learner to identify the target rectangle with as few training examples as possible. Remember that since the target concept is not initially known to the learner, your strategy cannot make use of it. Use the target concept $((4,6),(6,4))$ and the initial positive example $(5,5)$ to illustrate your strategy.

In this question, many of the solutions were not fully specified. When rounding is involved, there are two ways you can round, so please say which way you chose to round. Sometimes, it can even make a difference in the worst case cost. Also, many of the longer answers were harder to understand. You should try to be succinct and clear. Always describe the intuition behind your algorithm before giving any pseudo-code, it helps us understand what you are trying to do.

The best solution is a simple binary search in each of the four directions from the initial point (up, down, right and left). For each of the four directions, you always ask for the midpoint of the range of possible values for that particular edge of the target concept. When you have to round, always round away from the initial point.

As an example, consider the start point $(x,\ y) = (5,\ 5)$ and the target concept $((4,6),\ (6,4))$. We want to find this target concept, the corners of which we call $((x_L,\ y_T),\ (x_R,\ y_B))$, as shown in the diagram below.



First we find the right hand border x_R of the target concept. In our example, it can be at $x_R \in [5 \ldots 10]$. The midpoint is 7 or 8, but rounding away from the initial point means we ask for the label of the point $(8,5)$. $(8,5)$ is negative, so the right hand border of the target concept must be $x_R \in [5 \ldots 7]$. The midpoint is 6, so we ask for $(6,5)$. This is positive, so we know $x_R \in [6,7]$. The midpoint, rounded away from the start point is 7. So we ask for $(7,5)$, which is negative, and then we know that $x_R = 6$.

Similarly, to find the left hand border, we initially know $x_L \in [0 \ldots 5]$, with midpoint 2 or 3. Rounding away from the initial point, we ask for $(2, 5)$, which is negative. Then we know $x_L \in [3 \ldots 5]$, with midpoint 4, so we ask for $(4, 5)$, which is positive. Hence $x_L \in [3, 4]$. Finally, we ask for $(3, 5)$ which is negative. Therefore $x_L = 4$.

We do exactly the same sequence of operations for the top and bottom borders (except that for all the points we request, we use $x = 5$ and use the appropriate y value). This gives us $y_T = 6$ and $y_B = 4$ after three examples each.

Hence the target concept is $((4, 6),\ (6, 4))$.

Part d (15 POINTS)

For your strategy from Part c, what is the maximal number of training examples needed to identify any target concept? Illustrate your result by constructing the "worst case" concept and initital positive example, so that the number of training examples is maximal.

In this problem, often there was no justification why a given example might be worst case. We did not require a proof, but some sort of diagram or rough intuition why the example is worst case certainly helped, especially when using some unusual example as worst case. For many of the algorithms, the example given in part (c) was in fact a worst-case example. Another good one was with the initial point at (5,5) and the target concept being the entire hypothesis space, $((0, 10),\ (10, 0))$.

The best solution possible is always being able to learn the target concept with at most 12 examples, in addition to the one initially given. This is because with an optimally implemented binary search, you can select among n values in $\lceil log_2(n) \rceil$ steps.

If the initial point is in the middle of the hypothesis space, we need $\lceil log_2(5) \rceil = 3$ examples to pin down each of the four sides of the target concept, giving a total of $4 \times 3 = 12$ examples.

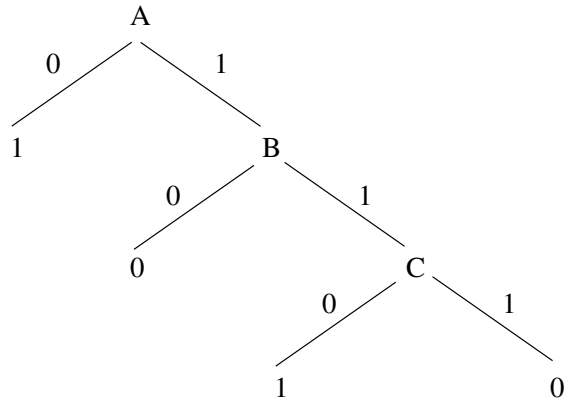
If we were to consider a starting point near the edge of the hypothesis space, say at $(1, 1)$, we need 1 example to check if each of $(0, 1)$ and $(1, 0)$ are positive or negative, and then $\lceil log_2(9) \rceil = 4$ examples for the two other sides. This gives a total of 10 examples, which is therefore not the worst case situation.

2 Decision Trees

Part a (10 POINTS)

What is a decision tree with as few nodes as possible that represents the boolean function $(\neg A \vee B) \wedge \neg(C \wedge A)$ over the boolean attributes A, B, and C?

$(\neg A \vee B) \wedge \neg(C \wedge A)$ is equivalent to $\neg A \vee (B \wedge \neg C)$ using De Morgan's Law and Associativity. The following tree represents this function:

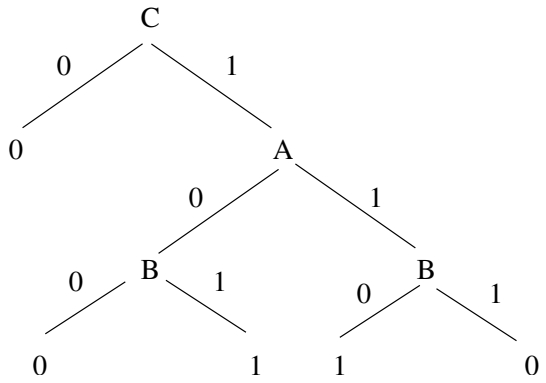


Since the minimal equivalent formula involves A, B and C, the smallest tree would have to have at least three nodes and therefore the given tree has as few nodes as possible.

Part b (10 POINTS)

What is a decision tree that represents the boolean function $(A \ XOR \ B) \wedge C$ over the boolean attributes A, B, and C?

$A \ XOR \ B$ is equivalent to $(\neg A \wedge B) \vee (A \wedge \neg B)$. So a possible tree representing this function is:



Part c (20 POINTS)

Grow the decision tree for predicting $SPAM$ for the following dataset using Information Gain as the splitting criterion. The attributes "nigeria", "viagra", and "learning" indicate whether that particular word occurs in the document. Show your calculations at each step.

nigeria	viagra	learning	SPAM
1	0	0	1
0	1	0	1
0	0	0	0
1	0	1	0
0	0	0	0
1	1	0	1
0	1	1	0
1	0	0	1
0	0	0	0
1	0	0	1

$$\begin{aligned} Entropy(Root) &= -(5/10) * log(5/10) - (5/10) * log(5/10) \\ &= 1 \end{aligned}$$

$$\begin{aligned} Gain(Root, Nigeria) &= 1 - (5/10) * (-(4/5) * log(4/5) - (1/5) * log(1/5)) - \\ &\quad (5/10) * (-(4/5) * log(4/5) - (1/5) * log(1/5)) \\ &= 0.278 \end{aligned}$$

$$\begin{aligned} Gain(Root, Viagra) &= 1 - (3/10) * (-(2/3) * log(2/3) - (1/3) * log(1/3)) - \\ &\quad (7/10) * (-(3/7) * log(3/7) - (4/7) * log(4/7)) \\ &= 0.035 \end{aligned}$$

$$\begin{aligned} Gain(Root, Learning) &= 1 - (2/10) * (-(0/2) * log(0/2) - (2/2) * log(2/2)) - \\ &\quad (8/10) * (-(5/8) * log(5/8) - (3/8) * log(3/8)) \\ &= 0.236 \end{aligned}$$

So $Nigeria$ is the appropriate choice of attribute for the root node.

$$\begin{aligned} Entropy(Nigeria = 1) &= -(4/5) * log(4/5) - (1/5) * log(1/5) \\ &= 0.722 \end{aligned}$$

$$\begin{aligned} Entropy(Nigeria = 0) &= -(4/5) * log(4/5) - (1/5) * log(1/5) \\ &= 0.722 \end{aligned}$$

$$\begin{aligned} Gain(Nigeria = 1, Learning) &= 0.722 - (1/5) * (-(0/1) * log(0/1) - (1/1) * log(1/1)) - \\ &\quad (4/5) * (-(4/4) * log(4/4) - (0/4) * log(0/4)) \\ &= 0.722 \end{aligned}$$

$$\begin{aligned} Gain(Nigeria = 1, Viagra) &= 0.722 - (1/5) * (-(1/1) * log(1/1) - (0/1) * log(0/1)) - \\ &\quad (4/5) * (-(3/4) * log(3/4) - (1/4) * log(1/4)) \\ &= 0.073 \end{aligned}$$

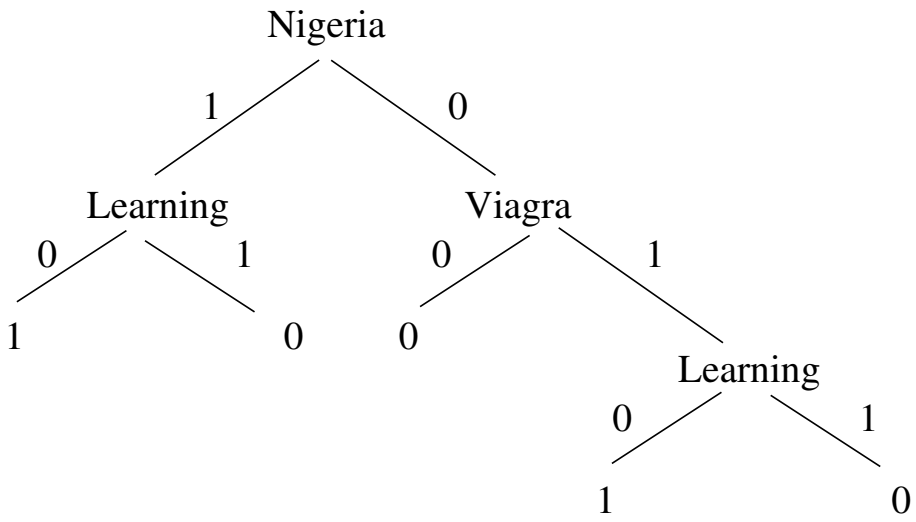
When $Nigeria = 1$, $Learning$ is the appropriate choice of attribute to split on. Both the children of $Learning$ are pure nodes so we stop here.

$$\begin{aligned} Gain(Nigeria = 0, Learning) &= 0.722 - (1/5) * (-(0/1) * log(0/1) - (1/1) * log(1/1)) - \\ &\quad (4/5) * (-(1/4) * log(1/4) - (3/4) * log(3/4)) \\ &= 0.073 \end{aligned}$$

$$\begin{aligned} Gain(Nigeria = 0, Viagra) &= 0.722 - (2/5) * (-(1/2) * log(1/2) - (1/2) * log(1/2)) - \\ &\quad (3/5) * (-(0/3) * log(0/3) - (3/3) * log(3/3)) \\ &= 0.322 \end{aligned}$$

When $Nigeria = 0$, $Viagra$ is the appropriate choice of attribute to split on. When $Nigeria = 0$ and $Viagra = 0$ the node is pure and so we stop there. When $Nigeria = 0$ and $Viagra = 1$, we split on the last remaining attribute i.e. on $Learning$.

So the resulting tree looks as follows:


Nigeria
1
0
Learning
Viagra
0
1
0
1
1
0
0
Learning
0
1
1
0