This homework is individual work. All assignments are due at the beginning of class on the due date. Assignments turned in late will drop 10 points for each period of 24 hours for which the assignment is late. In addition, no assignments will be accepted after the solutions have been made available.

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?

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.

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.

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.

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?

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?

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