There are six problems, each worth 20% of the exam. The problem with the lowest score will be treated as an extra credit problem, worth 1/4 of a regular problem.

You may use a calculator, but no notes or textbook. Accessing a phone or any other device besides a standalone calculator will be considered an act of academic dishonesty.

Do not turn this page until we give the signal.

Banner ID: __________________________
Problem 1 (Representations)
Consider the following training set with two classes in \( \mathbb{R}^2 \). Draw a decision tree that perfectly fits this training data. Make sure to completely specify all branches, internal nodes, and leaves.

Solution:

(Others are also correct.)
Problem 2 (Loss Functions)

Consider the 1-dimension hinge loss for a single training example \( x = 1 \) and \( y = 1 \) and a homogeneous halfspace with one weight \( w \in \mathbb{R} \):

\[
\ell_{\text{hinge}}(w) = \begin{cases} 
0 & \text{if } w \geq 1 \\
1 - w & \text{if } w < 1 
\end{cases} = \max\{0, 1 - w\}
\]

\( \ell_{\text{hinge}}(w) \) is a convex function that upper bounds another loss function called the ramp loss:

\[
\ell_{\text{ramp}}(w) = \begin{cases} 
0 & \text{if } w \geq 1 \\
1 - w & \text{if } 0 < w < 1 \\
1 & \text{if } w \leq 0 
\end{cases}
\]

a. Prove that \( \ell_{\text{ramp}}(w) \) is a non-convex function.

**Solution:** It is sufficient to show there exists \( w_1, w_2 \in \mathbb{R} \) and \( \alpha \in [0, 1] \) such that

\[
\ell_{\text{ramp}}(\alpha w_1 + (1 - \alpha)w_2) > \alpha \ell_{\text{ramp}}(w_1) + (1 - \alpha)\ell_{\text{ramp}}(w_2)
\]

Let \( w_1 = -1 \), \( w_2 = 1 \), and \( \alpha = 0.5 \). Then \( \ell_{\text{ramp}}(\alpha w_1 + (1 - \alpha)w_2) = 1 \) and \( \alpha \ell_{\text{ramp}}(w_1) + (1 - \alpha)\ell_{\text{ramp}}(w_2) = 0.5 \), completing the proof.

Arguing that the ramp loss has multiple local minima or that its epigraph is a nonconvex set are also acceptable.

b. When learning a halfspace over many training examples, what could be an advantage of using the hinge loss over the ramp loss, and vice versa, what could be an advantage of using the ramp loss over the hinge loss?

**Solution:** An advantage of the hinge loss is that it is convex, meaning that there will be one globally minimal value of the empirical risk over all weights.

An advantage of the ramp loss is that it bounds how much loss can be contributed to the total by a single example. It is therefore less sensitive to outlying data points that are far from the correct side of the decision boundary.

Other answers are also possible.
Problem 3 (Optimizers)
Consider the hypothesis class of two-dimensional thresholds, \( H = \{ h_{a,b} : a \in \mathbb{R} \text{ and } b \in \mathbb{R} \} \) where:

\[
h_{a,b}(x) = \begin{cases} 
1 & \text{if } x_1 \leq a \text{ and } x_2 \leq b \\
-1 & \text{otherwise}
\end{cases}
\]

and \( \mathcal{X} = \mathbb{R}^2 \) and \( \mathcal{Y} = \{-1, 1\} \).

Describe an algorithm for computing the ERM for this class in the realizable case. (You can assume 0-1 loss, although the solution will be the same for any reasonable loss function.) State the computational complexity of the algorithm in the context of a training data set of size \( m \).

Solution: Given this hypothesis class, finding the ERM entails calculating the values of \( a \) and \( b \).

Let \( a \) be the largest first element \( (x_1) \) of an example \( x \) with a label of 1, and let \( b \) be the largest second element \( (x_2) \) of an example with a label of 1. (If no examples have a label of 1, set \( a \) and \( b \) to minimum value that can be represented.)

\( O(m) \) to find the maximum values.
Problem 4 (Empirical and Expected Risk)

For this problem, we are looking for responses that both indicate your assessment as to a possible accuracy change and your understanding of the algorithm that led to this assessment. Answers should be two or three sentences long and focus on the relevant and important issue.

a. We have trained a logistic regression model (binary vector input, binary label, no regularization) on a data set. Then, we create a new data set that is identical to the original but it includes a new feature that is set uniformly at random, with no strong correlation to any of the other features or the label, and run the same learning algorithm again. What would you expect to happen to the training and testing losses of the new learned model?

**Solution:** We expect the training loss to decrease, because we have increased model complexity. Even though the new feature is not strongly correlated with the label, the additional feature corresponds to an additional parameter that the learning algorithm can use to fit the data.

However, we expect the testing loss to increase, because the learning algorithm will assign a non-zero weight to a feature that is random noise, i.e., it overfits at least slightly to that feature.

b. We have trained a logistic regression model (binary vector input, binary label, no regularization) on a data set. Then, we create a new data set that is identical to the original but includes a new attribute that is the Boolean negation of the label and run the same learning algorithm again. What would you expect to happen to the training and testing losses of the new learned model?

**Solution:** We expect the training loss of our algorithm to approach 0 as we train. This is because there is one feature that exactly corresponds to the opposite label, so its partial derivative is always negative (rigorous proof not needed). Therefore, its weight will decrease on each gradient update.

At the end of training, we expect our learned model to have loss approaching zero on both the training and test data, because it always predicts the opposite value of the feature as the label, and a very large, negative (possibly overflowing) weight for that feature.
Problem 5 (Model Selection)

Consider the following (partially labeled) model selection curve for boosted halfspace classifiers learned with AdaBoost:

![Model Selection Curve](image)

a. Describe the likely interpretation of the following parts of the above figure, based on the bias-complexity tradeoff. Include a specific statement of what that part of the figure represents, and provide a brief explanation justifying your interpretation.

The horizontal axis (with values 1 through 7):

**Solution:** The number of iterations of AdaBoost determines the complexity of the learned hypothesis. It is likely that this axis is the number of iterations because a model selection curve is used to choose a hyperparameter that trades off between bias and complexity.

Curve A (solid line):

**Solution:** It is likely the training error, since it continues to decrease as model complexity increases. Making a model more complex decreases, but does not increase, the training error.

Curve B (dotted line):

**Solution:** It is likely the validation error, because it starts to decrease as model complexity increases, but eventually grows again. Making a model more complex can increase the estimation error. If it increases more than the approximation error decreases, the error on the held-out validation set will grow.

b. If you were using the above model selection curve to choose a specific value on the horizontal axis to use for the corresponding task, which would you choose? Why?

**Solution:** 4 is a good choice because it minimizes the validation error.
Problem 6 (VC dimension)

Consider (again, see problem 3) the hypothesis class of two-dimensional thresholds, \( H = \{ h_{a,b} : a \in \mathbb{R} \text{ and } b \in \mathbb{R} \} \) where:

\[
h_{a,b}(x) = \begin{cases} 
1 & \text{if } x_1 \leq a \text{ and } x_2 \leq b \\
-1 & \text{otherwise}
\end{cases}
\]

and \( X = \mathbb{R}^2 \) and \( Y = \{-1, 1\} \).

What is the VC dimension of this hypothesis class? Provide a complete proof.

**Solution:** Recall that to show that the VC dimension of this hypothesis is \( k \), we need to show that

1. \( H \) shatters some set of size \( k \).
2. Every set of size \( k + 1 \) cannot be shattered by \( H \).

We will show that the VC dimension of \( H \) is 2:

1. \( H \) shatters some set of size 2. Consider the set of points \( X = \{(1,0), (0,1)\} \). \( H \) shatters \( X \) because for every mapping from \( X \) to \( Y \), we can find appropriate \( a \) and \( b \) values to correctly classify the points according to that mapping.

\[
\begin{array}{c|cc|c|c}
\hline
x & (1,0) & (0,1) & a & b \\
y & -1 & -1 & 0 & 0 \\
\hline
\end{array}
\]

\[
\begin{array}{c|cc|c|c}
\hline
x & (1,0) & (0,1) & a & b \\
y & 1 & -1 & 1 & 0 \\
\hline
\end{array}
\]

\[
\begin{array}{c|cc|c|c}
\hline
x & (1,0) & (0,1) & a & b \\
y & -1 & 1 & 0 & 1 \\
\hline
\end{array}
\]

\[
\begin{array}{c|cc|c|c}
\hline
x & (1,0) & (0,1) & a & b \\
y & 1 & 1 & 1 & 1 \\
\hline
\end{array}
\]

2. Every set of size 3 cannot be shattered by \( H \). Consider any three points \( x_1, x_2, x_3 \) all in \( \mathbb{R}^2 \). Without loss of generality assume that they are partially ordered along each dimension, so that \( x_{11} \leq x_{21} \) or \( x_{11} \leq x_{31} \), as well as \( x_{12} \leq x_{22} \) or \( x_{12} \leq x_{32} \).

Now, let \( x_1 \) be labeled \(-1\), and let \( x_2 \) and \( x_3 \) be labeled \(1\). To correctly classify \( x_2 \) and \( x_3 \), we must choose \( a \) and \( b \) such that \( x_{21} \leq a \) and \( x_{31} \leq a \), as well as \( x_{22} \leq b \) and \( x_{32} \leq b \). However, any such \( a \) and \( b \) will misclassify \( x_1 \), so we have shown that no set of size 3 can be shattered by \( H \).

Thus, we have shown that the VC dimension of \( H \) is 2.