CS 491 Midterm 2 Name ___________________________

Short Problems (20 points)

1. In the following statement, underline and indicate which part of this refers to recall, and which part refers to to precision. (4 points)
   "I Swear To Tell The Truth, The Whole Truth and Nothing But The Truth So Help Me God"

2. Let's say we have a single layer neural network. Our data has N samples and D features. We have computed the gradient of our Loss function with respect to our weights. In Big-O notation, what is the cost of one gradient descent update given the gradient? (4 points)

3. Suppose we take all the weights and biases in a network of perceptrons, and multiply them by a positive constant, $c > 0$. Does the behavior of the network change? Yes or no. Explain. (4 points)

4. AdaBoost will eventually reach zero training error, regardless of the type of weak classifier it uses, provided enough weak classifiers have been combined. True or False. Explain. (4 points)

5. Show for a matrix A and and eigenvector v that if $Av = \lambda v$ then $A^k v = \lambda^k v$. (4 points)

491

10 points

6. Consider the binary threshold neuron, $h = sign(w^T x)$ defined such that $h \in \{0, 1\}$, with no bias b or w_0. Consider the following set of four input features, x:

$$(1, 0, 0)^T,\ (0, 1, 0)^T,\ (0, 0, 1)^T,\ (1, 1, 1)^T$$

(a) Find a three-dimensional parameter vector w such that the neuron will have the output pattern $\{h\} = \{1, 1, 1, 1\}$ for the given four input features.

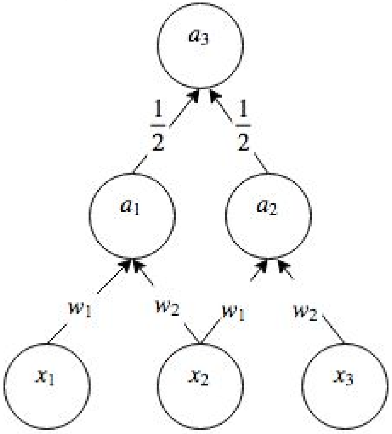
(b) Find a three-dimensional parameter vector w such that the neuron will have the output pattern $\{h\} = \{1, 1, 0, 0\}$ for the given four input features.

(c) Find an unrealizable output pattern $\{h\}$.

491

20 points

7. We depict a neural network below. There is weight sharing in the first level. This means that the two edges that are labeled w_1 are required to have the same value. If we change one, we change the other by the same amount. Similarly for w_2. There are no bias terms. Between the hidden layer and the output layer there are two edges labeled $\frac{1}{2}$. This is average pooling. These weights are fixed and never change. Backprop does not affect them. We also use a regression loss. So the network can be described with the equations:
$a_1 = w_1 x_1 + w_2 x_2$
$a_2 = w_1 x_2 + w_2 x_3$
$a_3 = \frac{1}{2} a_1 + \frac{1}{2} a_2$
$L = (y - a_3)^2$



We have one training example with $x_1 = 0$, $x_2 = 2$, $x_3 = 2$, $y = 2$. We initialize the network with $w_1 = -1$, $w_2 = 1$. For this training example:

(a) What are the values of a_1, a_2, a_3, and L?

(b) What is $\frac{\partial L}{\partial a_3}$?

(c) What is $\frac{\partial a_3}{\partial a_1}$?

(d) What is $\frac{\partial a_1}{\partial w_1}$?

(e) What is $\frac{\partial L}{\partial w_1}$?

491

10 points

8. Suppose we perform PCA on the four points with coordinates (x,y), (x,-y), (-x,y), (-x,-y).

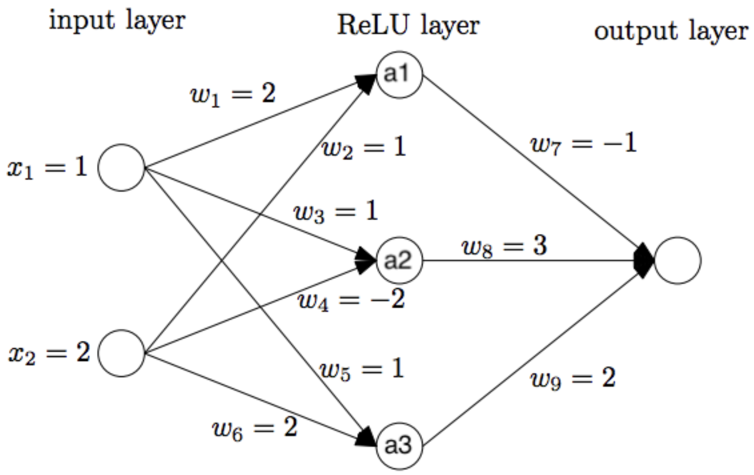
   (a) What is the covariance matrix for this data? (No variance scaling).

   (b) For what values of x and y will the principal component be $(1, 0)$?

491

20 points

9. Consider the neural network below. $L = (y - \hat{y})^2$ Assume the label for the input sample is $y = 3$ Compute the values of all weights w_i after performing a gradient descent update with learning rate 0.1. ReLU layer indicates that every neuron in that layer applies a ReLU activation.



491

20 points

10. In this problem, you will use Adaboost to learn a hidden function from this set of training examples. We will use two rounds of AdaBoost to learn a hypothesis for this data set. The Adaboost algorithm is provided on the back of this page for reference. Recall that in round number 1, AdaBoost chooses a weak learner that minimizes the weighted error ϵ. As weak learners, you will use axis parallel lines of the form

    - if $x_1 > a$, then $+1$ else -1 or
    - if $x_2 > b$, then $+1$ else -1, for some integers a, b.

    (either one of these two forms, not a combination of the two).

    Consider the following labeled data (x_1, x_2, y) where x_1 and x_2 are the attributes and y is the class variable:

sample	x_1	x_2	y
s_1	11	3	-1
s_2	10	1	-1
s_3	4	4	-1
s_4	12	10	+1
s_5	2	4	-1
s_6	10	5	+1
s_7	8	8	-1
s_8	6	5	+1
s_9	7	7	+1
s_{10}	7	8	+1

(a) The first step of AdaBoost is to create an initial data weight distribution D_1 (also called calculating the data weighting co-efficients). What are the initial weights given to data points s_4 and s_7 by the AdaBoost algorithm, respectively?

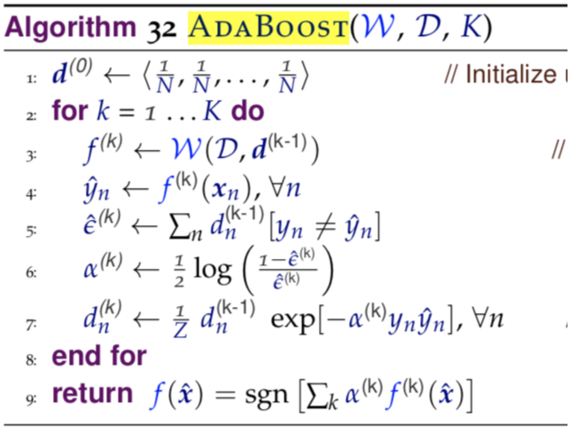
(b) Which of the following three hypotheses minimizes the weighted error in the first round of AdaBoost, using the distribution D_1 computed in the above question? Circle one. Justify your answer.

$x_2 > 9$ $\qquad$ $x_2 > 4$ $\qquad$ $x_2 > 7$

(c) What is the weighted error ϵ of the best classifier computed above in part (b)?

(d) Which of the following three hypotheses minimizes the weighted error in the second round of AdaBoost. Circle one. Justify your answer.

$x_2 > 9$ $\qquad$ $x_1 > 5$ $\qquad$ $x_2 > 7$



Algorithm 32 AdaBoost($\mathcal{W}$, $\mathcal{D}$, K)

```
$\boldsymbol{d}^{(0)} \leftarrow \langle \frac{1}{N}, \frac{1}{N}, \dots, \frac{1}{N} \rangle$    // Initialize
for $k = 1 \dots K$ do
    $f^{(k)} \leftarrow \mathcal{W}(\mathcal{D}, \boldsymbol{d}^{(k-1)})$    //
    $\hat{y}_n \leftarrow f^{(k)}(\boldsymbol{x}_n), \forall n$
    $\hat{\epsilon}^{(k)} \leftarrow \sum_n d_n^{(k-1)} [y_n \neq \hat{y}_n]$
    $\alpha^{(k)} \leftarrow \frac{1}{2} \log \left( \frac{1-\hat{\epsilon}^{(k)}}{\hat{\epsilon}^{(k)}} \right)$
    $d_n^{(k)} \leftarrow \frac{1}{Z} d_n^{(k-1)} \exp[-\alpha^{(k)} y_n \hat{y}_n], \forall n$
end for
return $f(\hat{\boldsymbol{x}}) = \mathrm{sgn}\left[\sum_k \alpha^{(k)} f^{(k)}(\hat{\boldsymbol{x}})\right]$
```