CS 422 Exam 2 Name ____________________________

Short Answer (10 Points)

1. Give two reasons why might we choose the ReLU activation function over a sigmoid activation function in a deep neural network? (5 points)

2. Explain the intuition behind Adaboost. Why do we use it? What does it give us?

Short Answer (10 Points)

3. List and define 5 hyperparameters for deep learning. (5 points)

4. True/False: PCA always produces orthogonal eigenvectors. Briefly explain. (5 points)

Adaboost (20 Points)

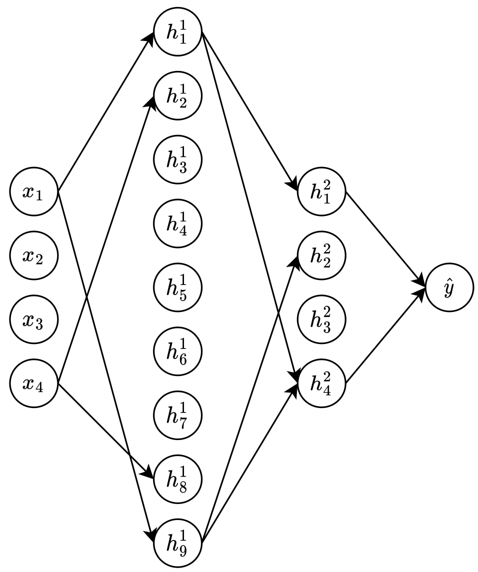
5. Assume you have four data points and you have four classifiers. The table below details which samples each classifier incorrectly classifies. Run two full iterations of adaboost, choosing the classifier with the lowest error at each iteration. Break ties by choosing h_i over h_k if $i < k$. In the table provided, indicate which classifier you chose, it's error rate and the weights for reach sample.

Classifier	Incorrect
h_1	s_1
h_2	s_1, s_2
h_3	s_2, s_3
h_4	s_4

Sample	Iter 0	Iter 1	Iter 2
s_1			
s_2			
s_3			
s_4			
h			
Error			

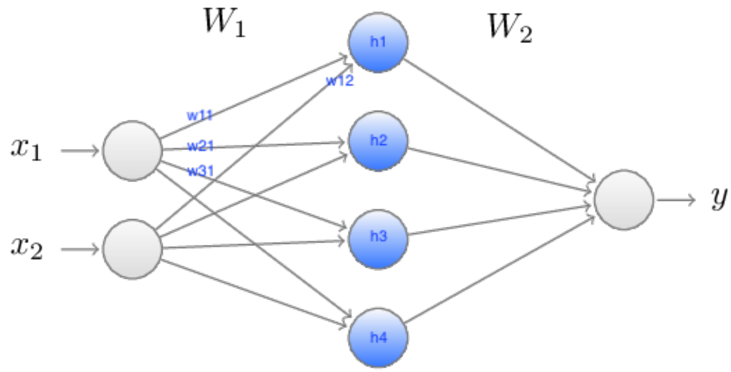
Deep Learning (10 Points)

6. Convert the following fully connected network to an equivalent convolutional neural network. That is, the networks should be equivalent in the number and application of weights to inputs, and hidden nodes. Assume the network is fully connected. I just didn't include all the connections because it takes too long to draw. Explain your choices.



Neural Networks (20 Points)

7. Given the following neural network, perform back propagation to update the following weights: w_{11} and w_{31}. Assume we are using a ReLU nonlinearity at the hidden nodes, a squared loss function $L(y, \hat{y}) = (y - \hat{y})^2$. Use the weights in the following table to start. Assume we have $x = (1, 1)$ as our input and $y = 1$ as our label. Also assume our learning rate is $\eta = 0.1$.



w_{11}	w_{21}	w_{31}	w_{41}	w_{12}	w_{22}	w_{32}	w_{42}	w_1^2	w_2^2	w_3^2	w_4^2
1	-1	2	-1	-2	1	1	2	-1	-2	1	0

CNNs (10 Points)

8. Assume you have a CNN. The input is of size 10×10 and you can apply filters each of size 3×3. How many layers would you need to get an output of size 2×2? How many parameters would this network have? Briefly explain.

PCA (20 Points)

9. Use PCA to find the principal components of the following data. No need to center or standardize the data. Project the following point onto the first principal component: $(3, 3)$. Project that same point onto both of the principal components.

Sample	x_1	x_2	y
s_1	0	1	1
s_2	1	0	1
s_3	1	1	1
s_4	2	2	-1
s_5	2	1	-1
s_6	1	2	-1

Equations

Adaboost

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

$\boldsymbol{d}^{(0)} \leftarrow \langle \frac{1}{N}, \frac{1}{N}, \ldots, \frac{1}{N} \rangle$
for $k = 1 \ldots K$ **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 $f(\hat{\boldsymbol{x}}) = \text{sgn}\left[\sum_k \alpha^{(k)} f^{(k)}(\hat{\boldsymbol{x}})\right]$

PCA

$$Cov = (X - \bar{X})(X - \bar{X})^T$$

$$|A - \lambda I| = 0$$

$$X^* = UX$$