CS 422/622 Exam 2 Name ____________________________

Short Answer (5 Points Each)

1. Is $f(x) = 2x$ a good choice for an activation function in a neural network? Why or why not?

2. Given the following eigenvectors and eigenvalues for a covariance matrix Z, what should we choose as our K value so that over 90% of the variance in the data is explained? Show your work or explain your answer.

$$v_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \; v_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \; v_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \; v_4 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

$\lambda_1 = 5$ $\lambda_2 = 4$ $\lambda_3 = 2$ $\lambda_4 = 1$

Short Answer (5 Points Each)

3. Give an example of a 3×3 kernel that could be used to detect vertical edges in an image. Briefly explain your choice.

4. Why is the ReLU used in place of the sigmoid as a non-linear activation function in modern neural networks?

Short Answer (5 Points Each)

5. Given a CNN layer whose input is $55 \times 55 \times 5$ and output is $49 \times 49 \times 12$. How many kernels are in this layer and what shape do they have? Assume no padding and a stride of 1.

6. How many parameters are there in a 3 layer neural network with 3 inputs, 4 nodes in the first hidden layer, 3 nodes in the second hidden layer and 2 output nodes? Show your work or explain your answer.

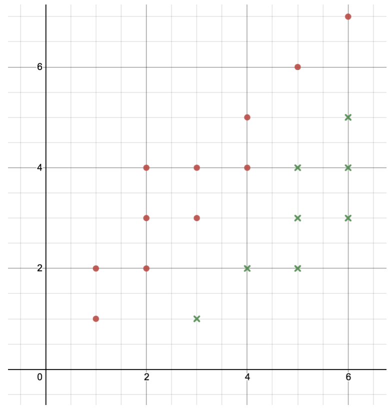
Short Answer (5 Points Each)

7. Why have CNNs replaced most other machine learning methods in image processing and computer vision problems?

8. True/False: PCA is a form of feature selection, keeping some features and throwing others away. Briefly explain.

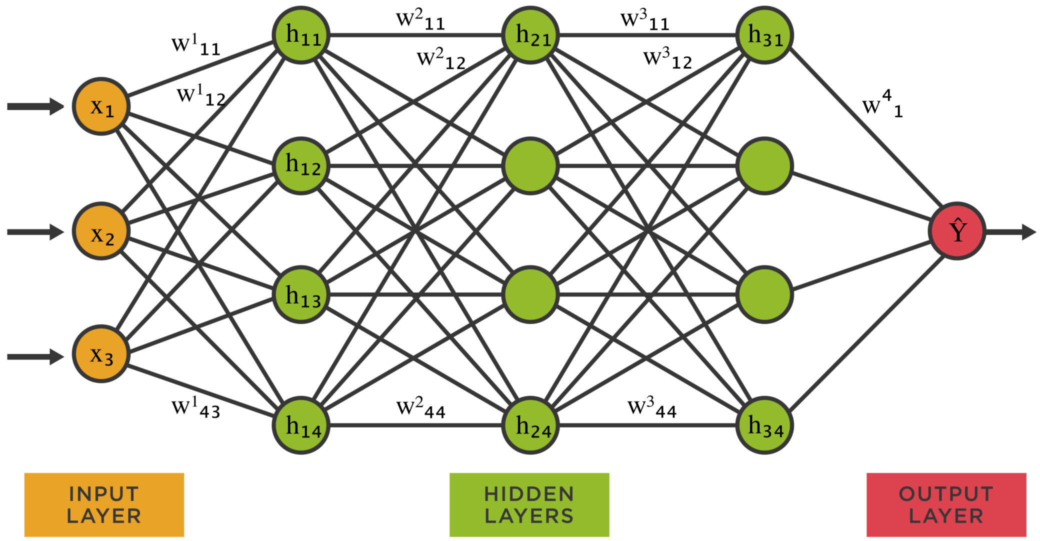
PCA (10 Points)

9. Given the following training data (circles indicate points with a positive label and Xs indicate points with a negative label), how could I use PCA to reduce the dimensionality of this data so that we could then use a linear classifier to perfectly classify it in 1D?



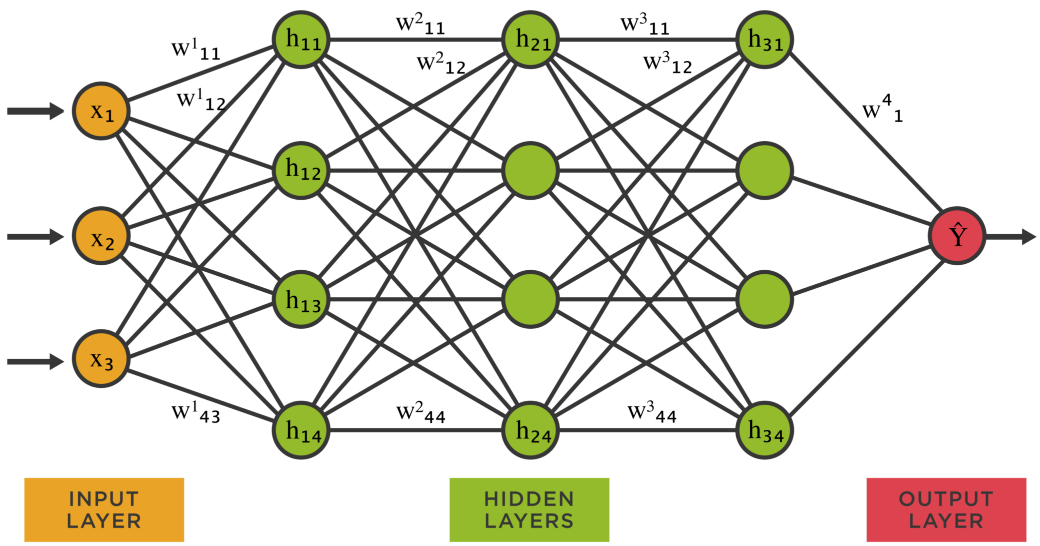
Neural Networks (10 Points)

10. Given the following network, give the update equation for w^3_{12}. Assume a nonlinear activation function of $f(x) = x^2$ at each hidden node and no nonlinear activation function at the output. Assume a loss function of $L(y, \hat{y}) = e^{-y\hat{y}}$.



Neural Networks (20 Points)

11. Given the following network, give the new value for w^2_{11} after one update using gradient descent. Assume a nonlinear activation function of $f(x) = x^2$ at each hidden node and no nonlinear activation function at the output. Use the following loss function: $L(y, \hat{y}) = (y - \hat{y})^2$. Assume all weights have a value of 1 and all biases have a value of 0. Use a learning rate of $\eta = 0.1$. Assume the input sample is $(0, 1, 0)$ and the label is 1025.



PCA (20 Points)

12. Given the following training data, where each row is a sample, find both principal components and project the data onto the first principal component. No need to standardize (divide by standard deviation) the data.

$$X = \begin{bmatrix} 0 & 0 \\ 2 & 1 \\ 4 & 2 \\ 6 & 3 \\ -1 & 2 \\ 1 & -2 \\ -2 & -1 \\ -4 & -2 \\ -6 & -3 \end{bmatrix}$$

Equations & Algorithms

PCA

Algorithm 37 PCA($\mathbf{D}$, K)

$\boldsymbol{\mu} \leftarrow \textsc{mean}(\mathbf{X})$ // compute data mean for centering
$\mathbf{D} \leftarrow \left(\mathbf{X} - \boldsymbol{\mu}\mathbf{1}^\top\right)^\top \left(\mathbf{X} - \boldsymbol{\mu}\mathbf{1}^\top\right)$ // compute covariance, $\mathbf{1}$ is a vector of ones
$\{\lambda_k, \boldsymbol{u}_k\} \leftarrow$ top K eigenvalues/eigenvectors of $\mathbf{D}$
return $(\mathbf{X} - \boldsymbol{\mu}\mathbf{1})\,\mathbf{U}$ // project data using $\mathbf{U}$