EEL5840/EEE4773 Fund. of Machine Learning
Summer 2022
Final Exam **Name:** ____________________
August 4, 2022
Time Limit: 2 hours UFID ____________________

- Write legibly
- There are a total of 9 questions for a total of 100 points
  - Some questions are worth more than other questions.
- **Closed-book, no computer, one-page formulas, calculator**
  - **Write your name in the formula sheet.**

Answer the questions in the spaces provided on the question sheets. If you run out of room for an answer, continue on the back of the page.

Closed-book, no computer, one-page formulas, calculator

Grade Table (for teacher use only)

Question:	1	2	3	4	5	6	7	8	9	Total
Points:	18	7	8	6	9	7	15	20	10	100
Score:										

1. (18 points) Answer the following questions regarding Fisher's Linear Discriminant Analysis (FLDA) and Principal Component Analysis (PCA).

   (a) (3 points) Consider two classes represented by two Gaussian distributions: G_1, with mean μ_1 and variance σ_1^2; and G_2, with mean μ_2 and variance σ_2^2. Using equations, define the within-class and between-class separation of G_1 and G_2.

   (b) (3 points) Provide an equation for the objective function used by FLDA on G_1 and G_2. What is the objective function trying to optimize?

(c) (5 points) What are the differences between FLDA and PCA? List at least 3 challenges of PCA and at least 3 challenges of FLDA. Justify your answers.

(d) (7 points) Suppose you want to project a D-dimensional data space to a 1-dimensional space. Show that PCA's direction of projection corresponds to the eigenvector (associated with largest eigenvalue) of the covariance matrix of the scaled feature matrix $\mathbf{X}$. Show all your work.

2. (7 points) Write down the pseudo-code for training the Perceptron algorithm.

3. (8 points) Answer the following questions regarding the soft-margin Support Vector Machine (SVM) classifier.

   (a) (4 points) Define the slack variable, ξ_n, in the soft-margin SVM. What is its role in the final solution?

   (b) (4 points) Suppose that you only want to penalize samples that are misclassified, propose a new slack variable and objective function to optimize this SVM.

4. (6 points) Suppose you have an MLP composed of one input layer with 10 neurons, followed by one hidden layer with 50 neurons, and finally one output layer with 3 neurons. All artificial neurons use the ReLU activation function, $\phi(x)$.

   (a) (1 point) What is the shape of the input matrix $\mathbf{X}$?

   (b) (1 point) What are the shapes of the hidden layer's weight vector W_h and its bias vector b_h?

   (c) (1 point) What are the shapes of the output layer's weight vector W_o and its bias vector b_o?

   (d) (1 point) What is the shape of the network's output matrix $\mathbf{Y}$?

   (e) (2 points) Write the equation that computes the network's output matrix $\mathbf{Y}$ as a function of $\mathbf{X}$, W_h, b_h, W_o, and b_o.

5. (9 points) Answer the following questions regarding an ANN architecture and justify your answers:

   (a) (3 points) How many neurons do you need in the output layer if you want to classify email into spam or ham? What activation function should you use in the output layer?

   (b) (3 points) If instead you want to tackle MNIST, how many neurons do you need in the output layer, and which activation function should you use?

(c) (3 points) What about for getting your network to predict housing prices? How many neurons do you need in the output layer, and which activation function should you use?

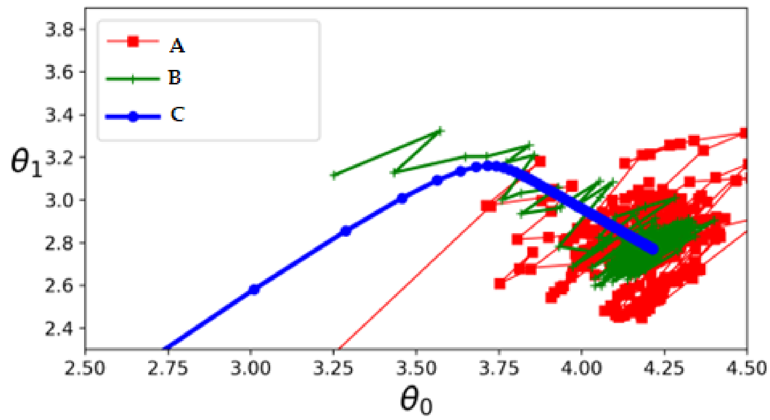
6. (7 points) Draw an Artificial Neural Network (ANN) that computes $A \oplus B$ (where $\oplus$ represents the XOR operation). Let all the bias terms be 0 and use the threshold activation function $\phi(x) = \begin{cases} 1, & x > 0 \\ 0, & x \leq 0 \end{cases}$. *Hint:* $A \oplus B = (A \wedge \neg B) \vee (\neg A \wedge B)$.

7. (15 points) Answer the following questions regarding training ANNs. Justify your answers.

   (a) (2 points) What is momentum optimization? Why is it useful? How is it integrated in backpropagation?

   (b) (2 points) What strategies can you use to avoid overfitting when training ANNs?

(c) (3 points) In the picture below, which curve (A, B or C) corresponds to mini-batch, online or batch learning?



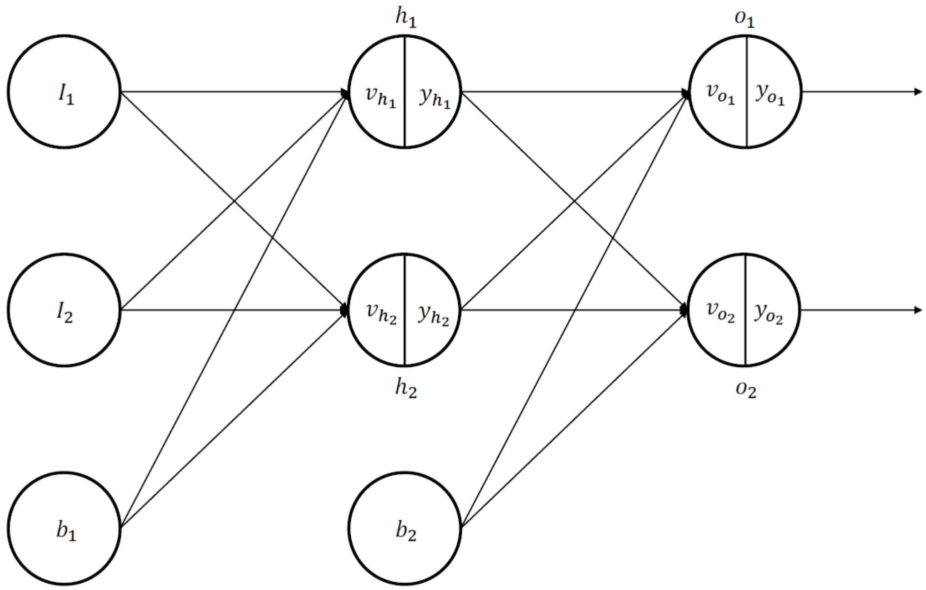
(d) (2 points) In a Convolutional Neural Network (CNN), why would you want to add a max pooling layer rather than a convolutional layer with the same stride? Justify your answer.

(e) (2 points) What strategies can you use to mitigate the vanishing/exploding gradient effects when training deep ANNs?

(f) (2 points) In a CNN, which filter size should you use if you want to capture color depth information only?

(g) (2 points) In practice, how would you determine whether to gather more data to train your classifier?

8. (20 points) Consider the following neural network:



with the initial weights and biases listed in the table below

weights/bias	connection	values
w_1	$I_1 \to h_1$	0.15
w_2	$I_2 \to h_1$	0.20
w_3	$I_1 \to h_2$	0.25
w_4	$I_2 \to h_2$	0.30
w_5	$h_1 \to o_1$	0.40
w_6	$h_2 \to o_1$	0.45
w_7	$h_1 \to o_2$	0.50
w_8	$h_2 \to o_2$	0.60
b_1		0.35
b_2		0.60

with all activation functions equal to the sigmoid function, $\phi(x) = \frac{1}{1+e^{-x}}$, and its derivative is, $\phi'(x) = \phi(x)(1 - \phi(x))$.

Consider the data point $x = [0.05, 0.10]^T$ with desired output vector $t = [0.01, 0.99]^T$. The objective function to be used to train this network is the squared error loss:

$$J = \frac{1}{2}\sum_{i=1}^{N}(t_i - y_i)^2$$

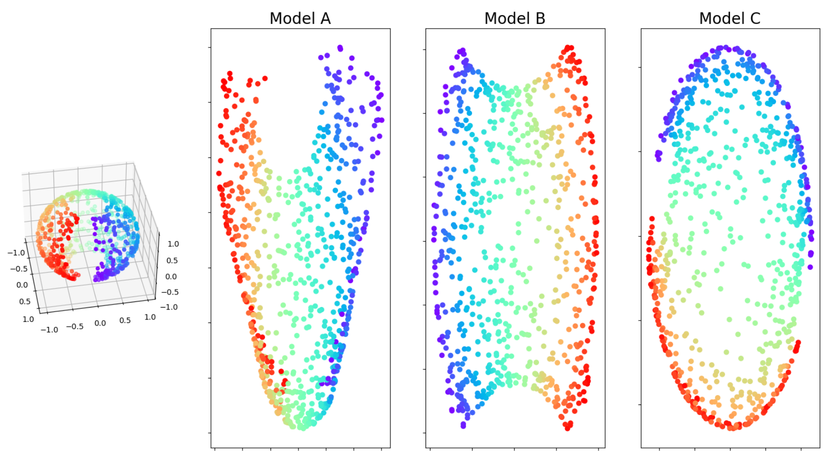
where t_i is the desired output value and y_i is the network output value. Answer the following questions:

(a) (5 points) Apply one forward pass using the point $x = [0.05, 0.10]^T$. What is the estimated output using this network?

(b) (7 points) Using the backpropagation algorithm, find the updated value for weight w_5 with a learning rate of $\eta = 0.1$.

(c) (8 points) Using the backpropagation algorithm, find the updated value for weight w_1 with a learning rate of $\eta = 0.1$.

9. (10 points) Consider the three-dimensional "open sphere" dataset displayed in the left-most plot below, as well as the performance of three different dimensionality reduction models (A, B and C):



Answer the following questions:

(a) (5 points) Which model performance (A, B or C) corresponds to Multi-Dimensional Scaling (MDS) with Euclidean distance, Locally Linear Embedding (LLE) and Isometric Mapping (ISOMAP)? Justify your answer.

(b) (5 points) Between MDS, LLE and ISOMAP, which algorithm is better equipped at preserving local structure and global structure of the manifold? Justify your answer.

HONOR STATEMENT

I understand that I am bound to uphold the honor code of the University of Florida. I have neither given nor received assistance on this examination. In addition, I did not use any outside materials on this exam other than the one page of formulas that was allowed.

Sign Your Name: ____

Write the Date: ____

Print Your Name: ____

Turn in your formula sheet with your exam!!!