Midterm Exam, Fall 2025
ESE 577

Instructor: Jorge Mendez-Mendez

Date: Thursday October 9th, 2025

Do not tear exam booklet apart!

- This is a closed book exam. One sheet (8 1/2 in. by 11 in.) of notes, front and back, are permitted. Calculators are not permitted.
- The total exam time is 80 minutes.
- The problems are not necessarily in any order of difficulty.
- Record all your answers in the places provided. If you run out of room for an answer, continue on a blank page and mark it clearly.
- If a question seems vague or under-specified to you, make an assumption, write it down, and solve the problem given your assumption.
- If you absolutely *have* to ask a question, come to the front.
- **Write your name on every piece of paper.**

Name: ______________________ SBU email: ______________________

Question	Points
1	21
2	16
3	18
4	21
5	24
Total:	100

Name: ____________________

Linear Regression

1. (21 points) We are given a small dataset and we would like to learn the parameters of a linear regressor hypothesis taking the form $h(x) = \theta^\top x + \theta_0$ for fitting the data.

   (a) Consider the following dataset $\mathcal{D}_1$ containing three data points (in feature-label pairs):

x	y
2	0
-3	-8
4	-2

   Suppose that we would like to minimize:

$$J_1(\theta, \theta_0; \mathcal{D}_1) = \frac{1}{n}\sum_{i=1}^{n}(\theta \cdot x^{(i)} + \theta_0 - y^{(i)})^2 \ ,$$

   where θ and θ_0 are both scalars.

   i. Call $\theta^{(1)^*}$ and $\theta_0^{(1)^*}$ the parameters that minimize J_1. Can we find $\theta^{(1)^*}$ and $\theta_0^{(1)^*}$ via the analytical solution formula? (No need for justification.)

   ○ Yes.

   ○ No.

   ii. Find a set of values $\theta^{(1)^*}$ and $\theta_0^{(1)^*}$ that minimize J_1.

(b) Suppose we add a second feature for each of the three datapoints in $\mathcal{D}_1$, to obtain the new dataset $\mathcal{D}_2$:

x_1	x_2	y
2	2.0001	0
-3	-3.0001	-8
4	4.0001	-2

Suppose that we would like to minimize:

$$J_2(\theta, \theta_0; \mathcal{D}_2) = \frac{1}{n}\sum_{i=1}^{n}(\theta^\top x^{(i)} + \theta_0 - y^{(i)})^2 \;\; ,$$

where θ_0 is still a scalar but θ is now a two-dimensional vector $\theta \in \mathbb{R}^2$.

i. Can we find ${\theta^{(2)}}^*$ and ${\theta_0^{(2)}}^*$ that minimize J_2 via the analytical solution formula? (No need for justification.)

○ Yes.

○ No.

ii. Compare ${\theta^{(1)}}^*$ and ${\theta^{(2)}}^*$. Which option below is true? Briefly justify your choice.

○ $\|{\theta^{(1)}}^*\| \ll \|{\theta^{(2)}}^*\|$

○ $\|{\theta^{(1)}}^*\| \gg \|{\theta^{(2)}}^*\|$

○ $\|{\theta^{(1)}}^*\| \approx \|{\theta^{(2)}}^*\|$

○ It depends

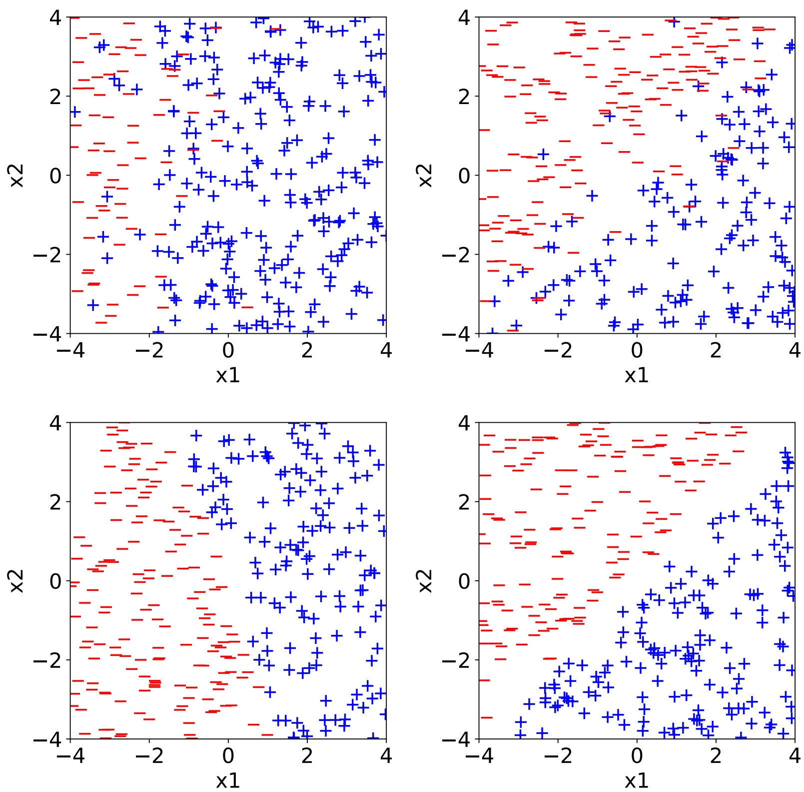
Justification:

Name: ______________________________

Logistic Mix-Up

2. (16 points) Below we show four different datasets in 2 dimensions. Points displayed as '–' have a negative label and points displayed as '+' have a positive label. We learned an unregularized linear logistic classifier for each of them, with a decision boundary given by $\theta^\top x + \theta_0 = 0$. But they all got mixed up! Please help us match each set of parameters to the dataset they came from, using the convention $\theta = \begin{bmatrix}\theta_1 & \theta_2\end{bmatrix}^\top$ and $x = \begin{bmatrix}x_1 & x_2\end{bmatrix}^\top$. (Each set of parameters is used exactly once.)

(a) (b)



(c) (d)

i. $\theta = \begin{bmatrix}10 & -10\end{bmatrix}^\top$, $\theta_0 = 0$: ◯ (a) ◯ (b) ◯ (c) ◯ (d)

ii. $\theta = \begin{bmatrix}1 & -1\end{bmatrix}^\top$, $\theta_0 = 0$: ◯ (a) ◯ (b) ◯ (c) ◯ (d)

iii. $\theta = \begin{bmatrix}1 & 0\end{bmatrix}^\top$, $\theta_0 = 2$: ◯ (a) ◯ (b) ◯ (c) ◯ (d)

iv. $\theta = \begin{bmatrix}9 & 3\end{bmatrix}^\top$, $\theta_0 = 0$: ◯ (a) ◯ (b) ◯ (c) ◯ (d)

Name: ______________________________

Regression Features

3. (18 points) Seawolves Game Day Predictions

You're using machine learning to predict the attendance at Stony Brook Seawolves basketball games. You have collected data from past games and need to encode the features properly before training a regression model.

For each of the following, suggest an appropriate feature encoding, and provide the dimension of the feature encoding.

(a) **Opponent:** The Seawolves regularly play against CAA teams, and there are 13 of them.

Encoding:

Encoding dimension:

(b) **Game importance:** People may be more likely to attend a game if it is more important. We identify three levels: "regular season", "playoffs", and "championship".

Encoding:

Encoding dimension:

(c) **Tip-off Time:** Games start at the hour sharp (e.g., 11 am, 12 pm, 7 pm) on certain hours of the day. Peak times are 11 am – 2 pm and 7 pm – 11 pm, with lower attendances in-between.

Encoding:

Encoding dimension:

Name: ______________________________

Gradient Descent

4. (21 points) We will do standard gradient descent iterations:

$$x^{(k+1)} = x^{(k)} - \eta \nabla f(x^{(k)}) \qquad (k = 0, 1, 2, \ldots) \ ,$$

on the following piecewise function:

$$f(x) = \begin{cases} (x+1)^2 & \text{if } x \leq 3 \\ (x-5)^2 + 12 & \text{if } x > 3 \end{cases} .$$

For the purposes of this question, assume that the gradient at $x = 3$ is $\nabla f(x)\Big|_{x=3} = 0$.

(a) Starting from the initial guess $x^{(0)} = -6$ and using a step size $\eta = 1$, what will be the values of x for the first three iterations of gradient descent, $x^{(1)}$, $x^{(2)}$, and $x^{(3)}$?

$x^{(1)} =$

$x^{(2)} =$

$x^{(3)} =$

(b) For each of the following statements, state whether the statement is true or false, and briefly (in one sentence) justify your choice.

i. For any function f, finding the point at which $\nabla f = 0$ is equivalent to finding the global minimum of that function

TRUE FALSE (Circle one)

Justification:

ii. For the piecewise function f in part (a), starting from $x^{(0)} = -6$ and using a step size $\eta = 1$—the same values from part (a)—gradient descent will converge to a minimum (local or global).

TRUE FALSE (Circle one)

Justification:

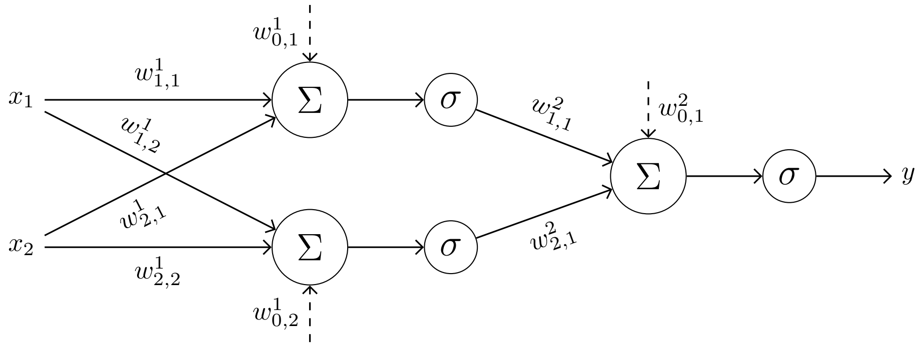
iii. For the piecewise function f in part (a), starting from $x^{(0)} = 8$ and using a step size $\eta = 1e{-}3$ fails to find the global minimum.

TRUE FALSE (Circle one)

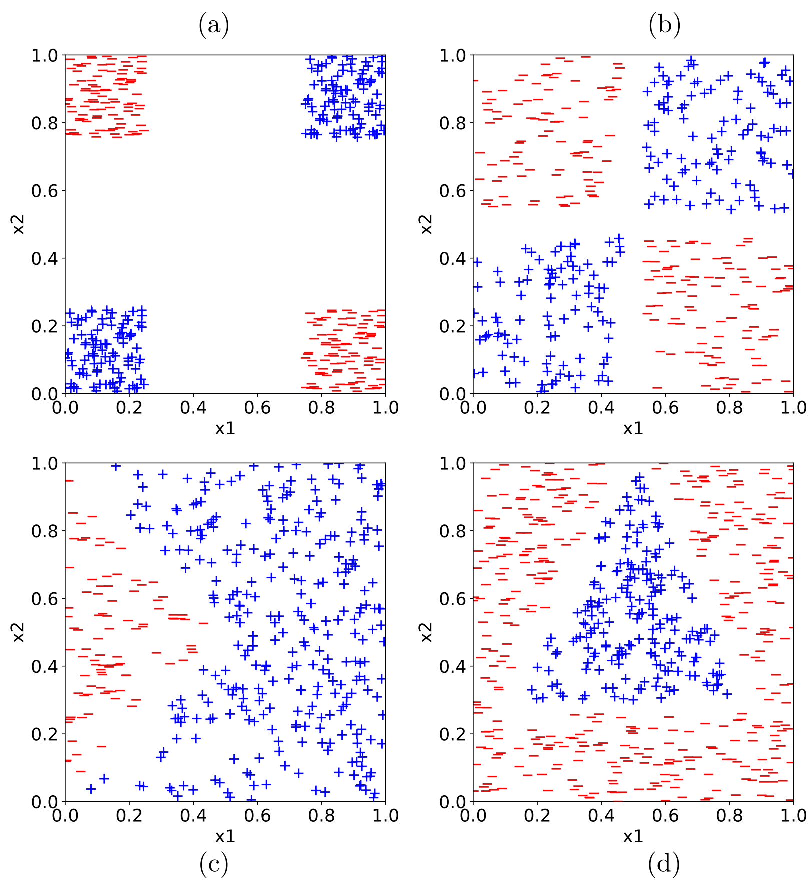
Justification:

Neural Networks

5. (24 points) We will explore what types of datasets can be perfectly separated by a simple neural network with two hidden units with sigmoid activation, as the one shown below. The two input features are x_1 and x_2, arrows into the summation nodes are annotated with the weights for the linear combination, biases are represented with dashed arrows, and σ represents the sigmoid activation.



Below we show four different datasets in 2 dimensions. Points displayed as '–' have a negative label and points displayed as '+' have a positive label. For each dataset, state whether it can be perfectly separated by the neural net. If it can, explain a possible transformation learned by each neuron that achieves perfect accuracy (e.g., "neuron one learns a linear separator that transforms points to the right of 0.8 to 1 and others to 0"). If not, briefly (in 1 sentence) explain why not.



Name: ____________________

(a)

Can be separated (circle one): Yes No
Justification:

(b)

Can be separated (circle one): Yes No
Justification:

(c)

Can be separated (circle one): Yes No
Justification:

(d)

Can be separated (circle one): Yes No
Justification:

Scratch paper Name: ____________________________________