Quiz-02

- Due Jan 26 at 11:59pm
- Points 10
- Questions 10
- Available Jan 24 at 6pm - Jan 26 at 11:59pm
- Time Limit None
- Allowed Attempts 3

Instructions

Learning in neural nets

This quiz covers topics from lectures 3 and 4, which cover the basics of learning in neural networks.

Topics in the quiz include those in the hidden slides in the slidedecks.

Take the Quiz Again

Attempt History

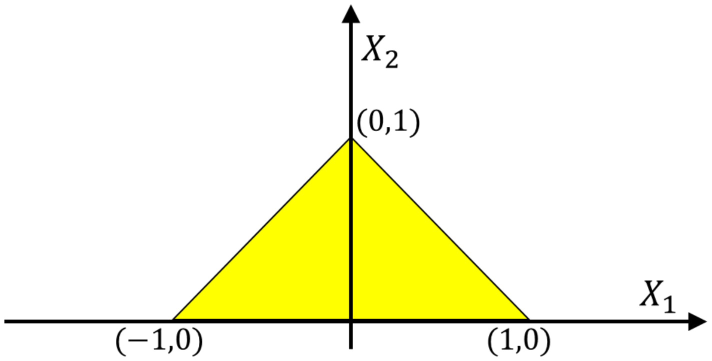
	Attempt	Time	Score
KEPT	Attempt 1	193 minutes	6 out of 10
LATEST	Attempt 2	258 minutes	5.5 out of 10
	Attempt 1	193 minutes	6 out of 10

Correct answers are hidden.

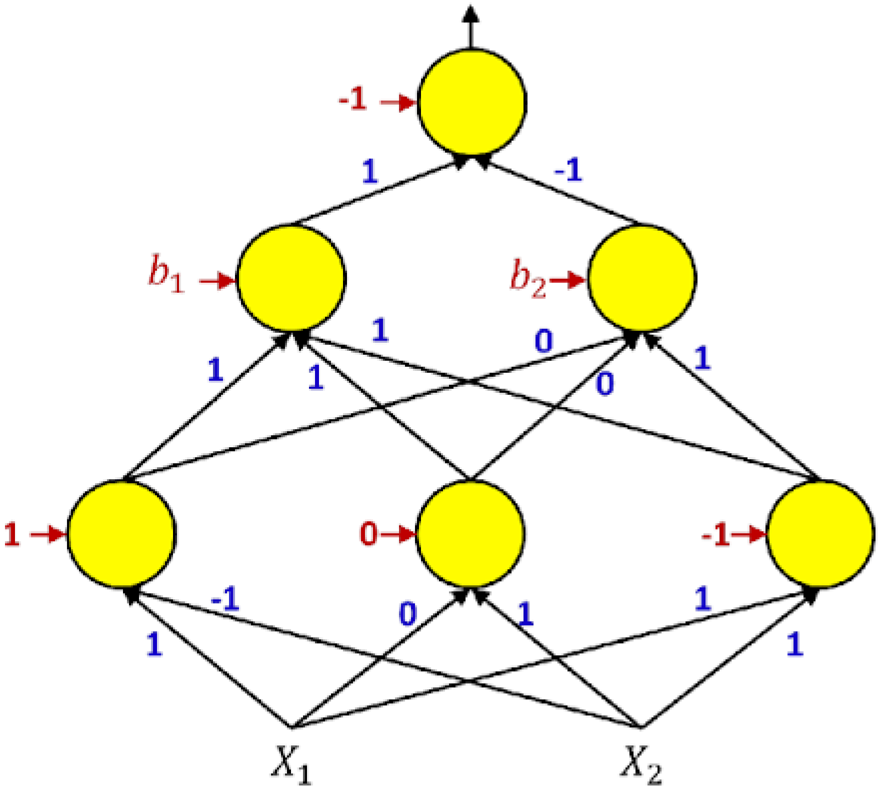
Score for this attempt: 5.5 out of 10
Submitted Jan 26 at 1:22pm
This attempt took 258 minutes.

PartialQuestion 1
0.5 / 1 pts

We want to build an MLP that composes the decision boundary shown in the figure below. The output of the MLP must be 1 in the yellow regions and 0 otherwise.



Consider the following suboptimal MLP with the given weights and biases:



Each perceptron of this MLP computes the following function:

$$y = \begin{cases} 1, & \sum_i weight_i input_i \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

The weights of the connections are shown in blue against the corresponding black arrows. The biases are shown in red for all but two perceptrons. What must the biases b1 and b2 be for the network to compute our target function perfectly? We require the biases to be integer values. Please give the value of b1 first and b2 second in the spaces provided below:

$b_1 =$ 0

$b_2 =$ -1

Hint: solve the equations. Lecture 2 Slide 83-93

Answer 1:

0

Answer 2:

-1

Question 2

1 / 1 pts

Gradient descent steps will always result in a decrease in the loss function we are minimizing .

Hint: See lec 4 slide 48 and 49

○ True

● False

The step size might be too large and the algorithm can reach a region where the value of the loss is more than that at the previous iteration.

Question 3

1 / 1 pts

For this question, please read these notes on the perceptron learning algorithm and select the correct options: **https://www.cse.iitb.ac.in/~shivaram/teaching/old/cs344+386-s2017/resources/classnote-1.pdf (https://www.cse.iitb.ac.in/~shivaram/teaching/old/cs344+386-s2017/resources/classnote-1.pdf)**

Hint: See lec 3, perceptron slides, and “logistic regression” slide

☑

We would like to change activation of the perceptron from the sign function to the sigmoid (σ) function to interpret it as a probability. For any input $\mathbf{x}^i$, we assume that $P(y^i = 1|\mathbf{x}^i) = \sigma(\mathbf{w} \cdot \mathbf{x}^i)$ and $P(y^i = -1|\mathbf{x}^i) = 1 - P(y^i = 1|\mathbf{x}^i)$. We then classify a point $\mathbf{x}^i$ as +1 if $P(y^i = 1|\mathbf{x}^i) \geq 0.5$ and as -1 otherwise. This sigmoid activated perceptron is still a linear classifier like the original perceptron.

☐

Suppose we have a set of n=100 points in d=3 dimensions which are linearly separable. Further assume that R=100 and γ=25. If we run the perceptron learning algorithm, then it will take **at least** 16 updates to converge.

☐

Since the algorithm takes at most $\frac{R^2}{\gamma^2}$ steps to converge, where R is the distance of the farthest point from the origin, if we scale down all the points by a constant factor 0<α<1, the new distance to the farthest point now reduces to αR. Thus, the algorithm would now take fewer steps to converge.

☐

Since the proof of convergence (Theorem 3) assumes that the points are linearly separable, it does not conclude anything about the non-linearly separable case. Therefore, in some cases, even if the points are not linearly-separable, the perceptron learning algorithm may still converge.

IncorrectQuestion 4

0 / 1 pts

(Select all that apply) Which of the following statements are true?

Hint: See slide Lec3 p79 and p89

MADALINE utilizes ADALINE to update neuron parameters

□

□ MADALINE is simply ADALINE, when it utilizes parallel computation

☑ ADALINE is used to train individual neurons, while MADALINE is used to train the entire network

☑

ADALINE uses a linear approximation to the perceptron that ignores the threshold activation. MADALINE, on the other hand, is greedy but exact.

Question 5

1 / 1 pts

(Select all that apply) For ADALINE, which of the following statements are true?

Hint: See slide Lec3 p80 - 83

☑ Has a linear decision boundary

□ Is equivalent to the generalized delta rule

☑ Moves weights in opposite direction of the gradient of the MSE (wrt the weights)

□ Is equivalent to the perceptron learning rule

☑ The calculated error is equivalent to that of a perceptron with identity activation

Both ADALAINE and Perceptron use variants of the delta rule, however ADALINE uses the output of the linear layer to calculate delta while perceptron uses the predicted output (after the threshold). For a perceptron with identity function, output of the linear layer will just be equal to output predicted output, hence in this case both will have the same error. For MSE, the derivative is negative and using the delta rule to update the weights will result in moving the weights in opposite direction of the gradient

Question 6

1 / 1 pts

You are performing gradient descent on the function $f(x) = x^2$. Currently, $x = 5$. Your step size is 0.1. What is the value of x after your next step?

Hint: See Lec 4 slides 40-43

4

Question 7

1 / 1 pts

A matrix is said to be positive definite if all of its Eigenvalues are positive. If some are zero, but the rest are positive, it is positive semi-definite. Similarly, the matrix is negative definite if all Eigen values are negative. If some are negative, but the rest are zero, it is negative semidefinite. If it has both positive and negative Eigenvalues, it is “indefinite”.

An N-dimensional function has an NxN Hessian at any point. The Eigenvalues indicate the curvature of the function along the directions represented by the corresponding Eigenvectors of the Hessian. Negative Eigen values indicate that the function curves down, positive Eigenvalues show it curves up, and 0 Eigenvalues indicate flatness.

(Select the correct answer) The Hessian of the function $f(x_1, x_2, x_3) = x_1^2 + 3x_2^2 + 2x_3^2$ at the point $(-1, \sqrt{2}, -1)$:

Hint: See lec 4, slide 19, 33-34, and rewatch that portion of the lecture. You will have to work out the Hessian and compute its Eigenvalues.

● Positive definite

Hessian: [[2,0,0], [0,6,0], [0,0,4]] and eigenvalues: 2, 6, 4

○ Positive semidefinite

○ Indefinite

○ Negative definite

○ Negative semidefinite

IncorrectQuestion 8

0 / 1 pts

Suppose Alice wants to meet Bob for a secret meeting. Because it is a secret meeting, Bob didn’t tell Alice the exact location where the meeting would take place. He, however, told her where to start her journey from and gave her directions to the meeting point. Unfortunately, Alice forgot the directions he gave to her. But she knows that the meeting would take place at the top of a hill close to her starting location.

Suppose the elevation of the ground that she is standing on is given by the equation $z = 20 + x^2 + y^2 - 10\cos(2\pi x) - 10\cos(2\pi y)$ where x, y are the 2-D coordinates and z is the elevation.

Alice decides to apply what she learned about function optimization in her DL class to go to the secret location. She decides to modify the gradient descent algorithm and walks in the direction of the fastest increase in elevation (instead of going opposite to the direction of fastest increase), hoping to reach the top of the hill eventually. Suppose she starts at the point **(3.1, -2.2)** and uses a step size (learning rate) of 0.001. At what point would she end up after taking 100 such steps? Truncate your answer to 1 digit after the decimal point.

Hint: See Lec 4 slides 40-43. The answer will require simulation.

$x =$ -1.5

$y =$ -0.5

Answer 1:

-1.5

Answer 2:

-0.5

IncorrectQuestion 9

0 / 1 pts

Which of the following statements are true, according to lecture 4? **(select all that apply)**

Hints: Lecture 4 discussion on derivatives (Slides 5-7), lecture 4 discussion on divergence, and lec 4 – individual neurons (Slides 64-65).

☑ The derivative of a function $y = f(x)$ with respect to its input x is the ratio $\frac{dy}{dx}$ of small increments in the output that result from small increments of the input.

☑ The derivative of a function $f(x)$ with respect to a variable z tells you how much minor perturbations of z perturbs $f(x)$

☑ The derivative $\nabla_x f$ of a function $f(x)$ of a vector argument x, with respect to x, is the same as the gradient of $f(x)$ with respect to x.

☑ Making the activation functions of the neurons differentiable enables us to determine how much small perturbations of network parameters influence the number of training data instances that are misclassified, and so helps us determine how to modify the parameters to reduce this number.

☑ It is necessary for both the activations and the divergence function that quantifies the error in the output of the network to be differentiable functions in the function minimization approach to learning network parameters.

□ The actual objective of training is to minimize the average error on the training data instances.

If you got any of these wrong, please watch the portions of the lecture corresponding to the hints.

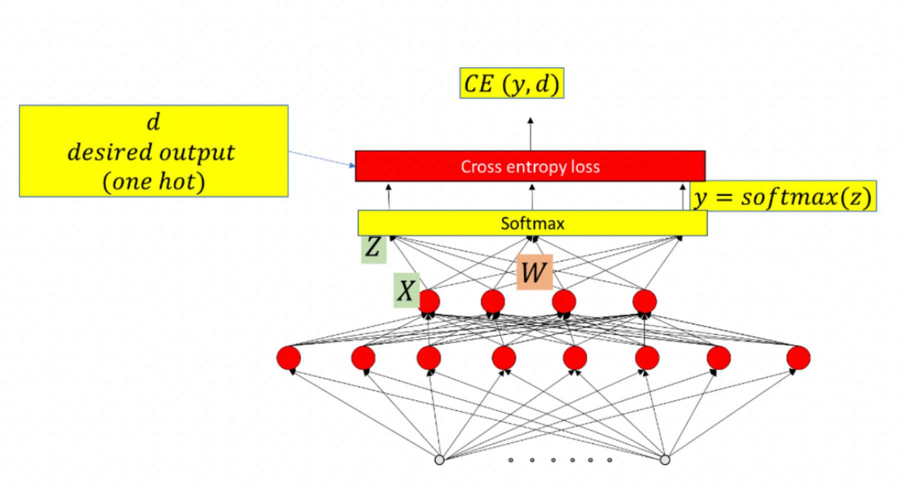
IncorrectQuestion 10

0 / 1 pts

A three-class classification neural network computes a 4-dimensional embedding X at the penultimate layer, just before the final classification layer, as shown in the figure. This is followed by a weight matrix W which computes an affine value Z (also called a logit) to which a softmax activation is applied to compute class probabilities.

Assuming row vector notation, as in Python, let the embedding vector $X = [1\ 2\ 3\ 4]$. Let the weight matrix $W = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}$.

The correct class (the true label) for this instance is class 1 (assuming classes are numbers 1 2 and 3). What is the gradient of the cross-entropy loss for this instance with respect to the affine (logit) vector Z?



○ [1 0 0]

○ [-1 0.33 0.33]

○ [1 0.33 0.33]

○ [1 0.27 0.73]

○ [-1 0.27 0.73]

● [0 0.27 0.73]

Quiz Score: 5.5 out of 10