

# Quiz-02

- Due Sep 8 at 11:59pm
- Points 10
- Questions 10
- Available Sep 6 at 6pm - Sep 8 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
LATEST	<a href="#">Attempt 1</a>	137 minutes	6 out of 10

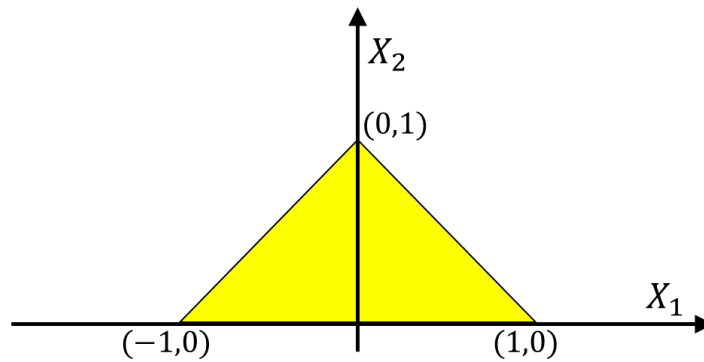
❗ Correct answers are hidden.

Score for this attempt: 6 out of 10  
Submitted Sep 7 at 5:59pm  
This attempt took 137 minutes.

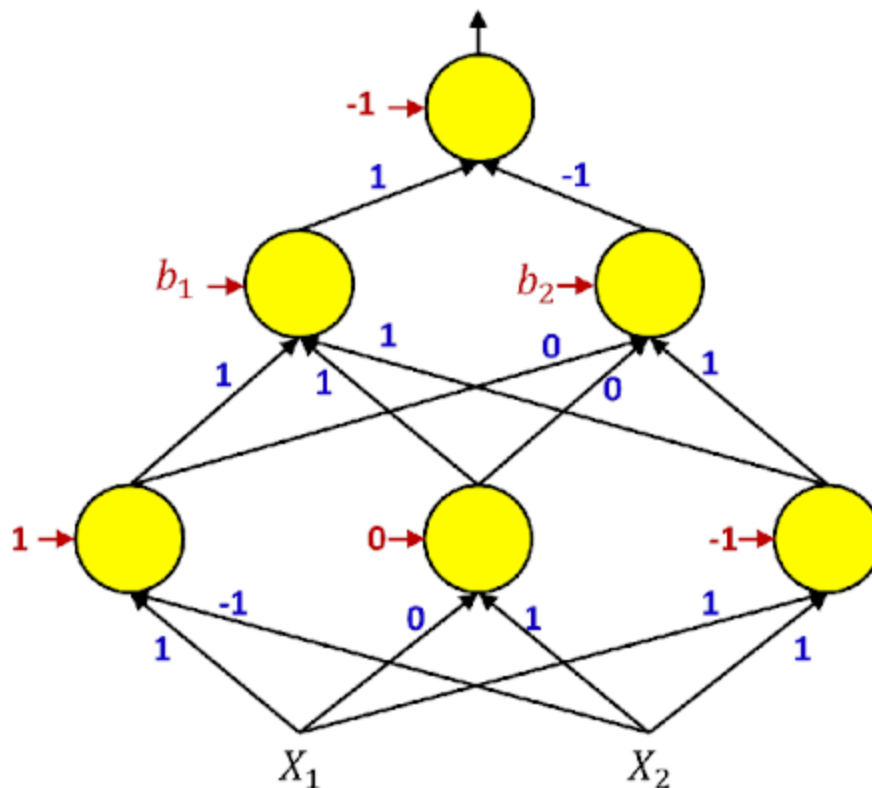


IncorrectQuestion 1  
0 / 1 pts

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:

$$y = \begin{cases} 1, & \sum_i \text{weight}_i \text{input}_i \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

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  $b_1$  and  $b_2$  be for the network to compute our target function perfectly? We require the biases to be integer values. Please give the value of  $b_1$  first and  $b_2$  second in the spaces provided below:

$b_1 =$

$$b_2 = -2$$

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

**Answer 1:**

-3

**Answer 2:**

-2



PartialQuestion 2

0.67 / 1 pts

**(Select all that apply)** Which of the following is true of the gradient of a function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$ , computed at any point

**Hint: See slide Lec4 p 8-10**

☒ The length of the gradient is indicative of the actual rate of change of the function, in the direction of the gradient

☐ You can always compute the gradient of any function at any location



The gradient is a vector composed of the partial derivatives of the scalar output of a function with respect to the components of its vector input


☒ The gradient is a vector that points in the direction of fastest increase of the function at that point

☐ The gradient is a vector that points in the direction of fastest decrease of the function at that point



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**



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  $0 < \alpha < 1$ , the new distance to the farthest point now reduces to  $\alpha R$ . 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.



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



We would like to change activation of the perceptron from the sign function to the sigmoid ( $\sigma$ ) 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.



PartialQuestion 4

0.33 / 1 pts

**(Select all that apply)** As stated in the lecture, why do we change the activation function from the threshold function?

**Hint: See Lec3, slides 93-100**

- ☒ Because we desire non-zero derivatives over contiguous regions of the input space
- ☒ Because we want to be able to determine how minor tweaks in parameters affect the empirical error
- ☒ Because the threshold function is never differentiable.

The threshold function is differentiable almost everywhere, but the derivative is 0, so it does not provide information about whether a change was for the better or worse.

- ☐ Because it helps us use the Gradient Descent technique



PartialQuestion 5

0.5 / 1 pts

**(Select all that apply)** How does ADALINE resolve the non-differentiability of the threshold activation?

**Hint: See slide Lec3 p79 - 91**

- ☐ It computes the squared error between the output of the perceptron and the target output, instead of counting errors.
- ☒ It tries to minimize the error between the desired binary output and the affine combination of inputs before the threshold activation is applied.
- ☐ It ignores the threshold activation during training, and only applies it during testing.
- ☐ It uses a differentiable sigmoidal approximation to the threshold function during learning, but uses the hard threshold activation subsequently when operating on test data.



### Question 6

1 / 1 pts

**(Select all that apply)** Which of the following procedures will give us the minimum point of a function  $f(x)$  that is twice differentiable and defined over the reals?

**Hint: See Lec4 slide 28**

- ☐ Computing the second derivative  $f''(x)$  and find an  $x$  where  $f''(x) > 0$  and  $f'(x) > 0$
- ☐ Computing the second derivative  $f''(x)$  and find an  $x$  where  $f''(x) = 0$  and  $f'(x) = 0$
- ☒ Computing the second derivative  $f''(x)$  and find an  $x$  where  $f''(x) > 0$  and  $f'(x) = 0$
- ☐ Computing the second derivative  $f''(x)$  and find an  $x$  where  $f''(x) < 0$  and  $f'(x) = 0$



### 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

$f(x_1, x_2, x_3) = x_1^2 x_2 + x_2^2 x_3 + x_3^3 + 2x_1 x_3 + x_2 x_3 + x_1 x_2$  at the point  $(-1, -1, -1)$  is :

**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 semidefinite
- ☒ Negative definite

Hessian:  $\begin{bmatrix} -2 & -1 & 2 \\ -1 & -2 & -1 \\ 2 & -1 & -6 \end{bmatrix}$  and eigenvalues: -6.8922, -2.9083, -0.1996

- ☐ Negative semidefinite
- ☐ Positive definite
- ☐ Indefinite



### Question 8

1 / 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 **(-1.8, -0.2)** 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.

$x =$

$y =$

**Answer 1:**

-1.5

**Answer 2:**

-0.5



PartialQuestion 9

0.5 / 1 pts

Which of the following statements are true **(select all that are true)**

Hint : Please watch the portions of the lecture where we explain linear vs affine. Also

[https://en.wikipedia.org/wiki/Linear\\_function](https://en.wikipedia.org/wiki/Linear_function)  [https://en.wikipedia.org/wiki/Linear\\_function](https://en.wikipedia.org/wiki/Linear_function) is useful. Also check Lecture 2, slides 13-14

☐ A function  $f(x)$  is said to be affine if and only if  $f(ax + by) = af(x) + bf(y)$  for all two scalar constants  $a$  and  $b$ .



A linear function represents the equation of a hyperplane that passes through the origin, whereas an affine function represents the equation of a hyperplane that may not pass through the origin.

A perceptron applies an activation to an affine combination of the inputs.

☐

☐ A function  $f(x)$  of vector  $x$  is said to be linear if  $f(Ax + By) = Af(x) + Bf(y)$  for all (square) constant matrices  $A$  and  $B$ .

☐ An affine function is linear, but not all linear functions are affine.

☒ A function  $f(x)$  is said to be linear if  $f(ax + by) = af(x) + bf(y)$  for all scalar constants  $a$  and  $b$ .

☐

A linear relation represents the equation of a hyperplane that is planar everywhere, whereas the surface represented by an affine relation can have non-linear behavior outside the domain of interest.

☒ A function  $f(x)$  is said to be affine if we can construct a  $g(x) = f(x) - c$  such that  $g(x)$  is linear, for some  $c$ .

☒ A perceptron applies an activation function to a linear combination of the inputs

A linear function is a function that has the property

$$f(ax + by) = af(x) + bf(y)$$

They have the property that  $f(0) = 0$  (because  $f(0) = f(0 \cdot x) = 0 \cdot f(x) = 0$ , regardless of  $x$ ).

$f(x) = 0$  will be a hyperplane that goes through origin.

An affine function is a linear function plus a bias. An affine  $f(x)$  has the property that  $f(x) = 0$  is a hyperplane that does not go through origin.



IncorrectQuestion 10

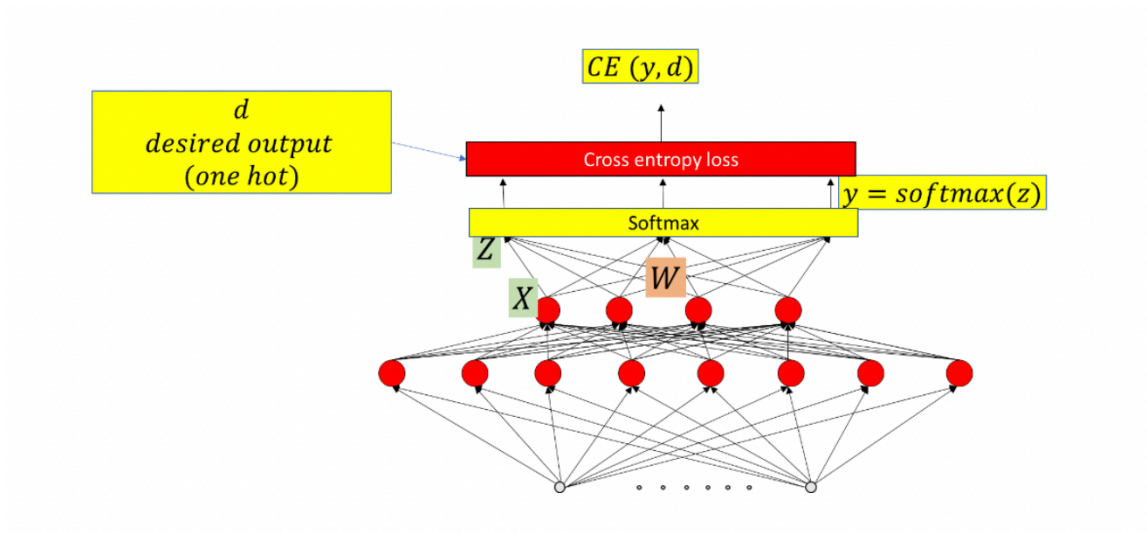
0 / 1 pts

A three-class classification neural network computes a 4-dimensional embedding  $\mathbf{X}$  at the penultimate layer, just before the final classification layer, as shown in the figure. This is followed by a weight matrix  $\mathbf{W}$  which computes an affine value  $\mathbf{Z}$  (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  $\mathbf{X} = [1 \ 2 \ 3 \ 4]$ . Let the weight

$$\text{matrix } \mathbf{W} = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}.$$

The correct class (the true label) for this instance is class 1 (assuming classes are numbers 1 2 and 3). What is the cross-entropy loss for this instance? Please provide the answer in the format X.XX (truncating to two decimals without rounding). Recall that the cross-entropy loss uses the natural log and not log base 10.



1.099

Quiz Score: 6 out of 10