Quiz-02

Started: Jan 26 at 5:51am

Quiz Instructions

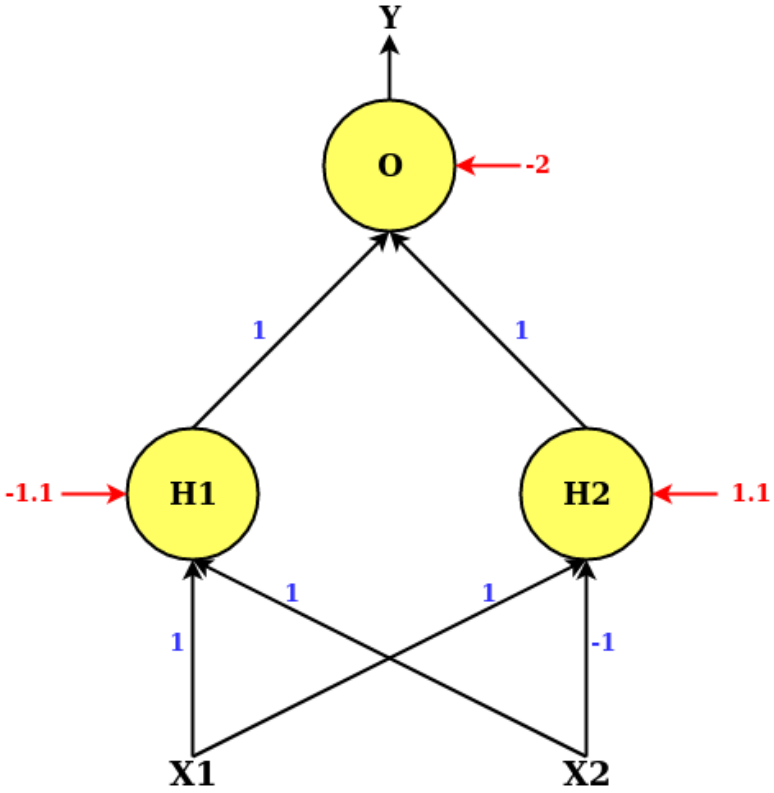
Learning in neural nets

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.

Question 1 1 pts

Consider the following MLP and the given parameters:



X1 and **X2** are the inputs to the network. **Y** is the output of the network. **H1** and **H2** are the hidden neurons and **O** is the output neuron. The weights of the connections are shown in blue against the corresponding black arrows. The biases are shown in red. Each neuron uses the threshold activation function:

$$\phi(z) = \begin{cases} 1, \text{ if } z >= 0 \\ 0, \text{otherwise} \end{cases}$$

If the inputs to the network are **X1**=0 and **X2**=0 and the desired output is **d**=0, which of the neurons will be updated first assuming that the MADALINE update rule is used?

Hint: See hidden ADALINE and MADALINE slides. Lecture 3 slide 86-92

- ○ The output neuron O
- ○ The first hidden neuron H1
- ○ All the neurons would be updated in the first step
- ◉ None of the neurons would be updated
- ○ The second hidden neuron H2

Question 2 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

Question 3 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)](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.
- ☑ 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.
- ☐ 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.

- [x] 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.

Question 4 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

- [x] Because we want to be able to determine how minor tweaks in parameters affect the empirical error
- [ ] Because it helps us use the Gradient Descent technique
- [ ] Because we desire non-zero derivatives over contiguous regions of the input space
- [x] Because the threshold function is never differentiable.

Question 5 1 pts

(Select the correct answer) For MADALINE, which of the following are true?

Hint: See slide Lec3 ADALINE and MADALINE slides

lect3: slide 86-92

- ○ Computes the gradient with respect to all the weights in the network
- ○ Updates the weights of at least one neuron with every training example
- ◉ Employs a chain rule to compute the derivatives of the error with respect to weights
- ○ It greedily selects a node with the minimum confidence (the affine combination closest to the threshold) and flips it

Question 6 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

4.5

Question 7 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.

- (●) Indefinite
- ( ) Negative semidefinite
- ( ) Positive definite
- ( ) Negative definite
- ( ) Positive semidefinite

Question 8 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.

$x =$ -1.5

$y =$ -0.5

Question 9 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).

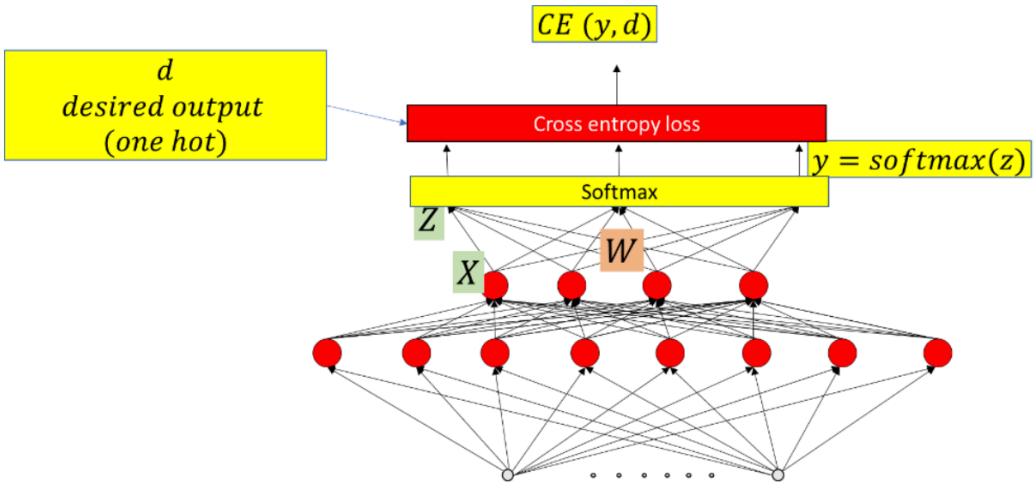
- [x] The derivative of a function $y = f(x)$ with respect to its input x is the ratio $\frac{dy}{dx}$ of small increments in the output that result from small increments of the input.
- [x] The actual objective of training is to minimize the average error on the training data instances.
- [x] It is 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.
- [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.
- [x] 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.

Question 10 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 $W = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}$.

What is the probability computed for class 3 by the network (assuming classes are number 1 2 and 3)? Please provide the answer in the format X.XX (two decimals), rounding up the second decimal value if necessary.



0.27

Quiz saved at 8:49am Submit Quiz