Quiz-02

Started: Jan 26 at 1:25pm

Quiz Instructions

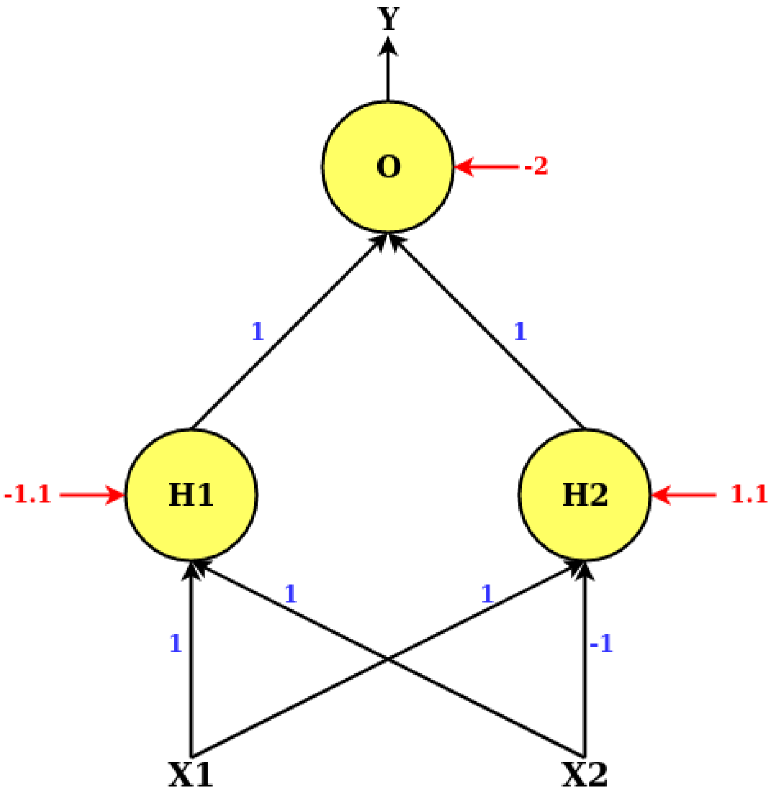
Learning in neural nets

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

○ All the neurons would be updated in the first step

● None of the neurons would be updated

○ The output neuron O

○ The first hidden neuron H1

○ The second hidden neuron H2

Question 2 1 pts

(Select all that apply) In order for our NN to represent a specific function, *in practice* we will actually try to:

Hint: See slide Lec3 p130-135

□ Maximize the network’s parameters

□ Minimize the weights and biases

☑ Adjust network parameters such that the network’s output matches the desired output as closely as possible, on the training instances

☑ Minimize the empirical error on the training data

Question 3 1 pts

(Select all that apply) Networks of perceptrons with threshold activations are hard to train because:

Hint: See slide Lec3 p97-p99

□ Threshold activations are inadequate to approximate most functions.

□ The computational complexity of identifying the appropriate labels for each of the training instances for each of the hidden perceptrons may be exponential in the number of training instances.

☑ The training data usually only provides labels for the entire network, and not for individual neurons in the network.

☑ We cannot generally get any indication of whether increasing any particular parameter will increase or decrease the overall error.

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

☐ Because the threshold function is never differentiable.

☑ Because we desire non-zero derivatives over contiguous regions of the input space

☑ Because it helps us use the Gradient Descent technique

☑ Because we want to be able to determine how minor tweaks in parameters affect the empirical error

Question 5 1 pts

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

Hint: Neural networks are universal function approximators. Sigmoid activation is a continuous function and it ranges from 0 to 1. No non-zero gradients pass through the threshold activations. Any addition of linear functions is still a linear function.

☑ A network with linear activation functions is equivalent to a single perceptron with linear activation

☑ The output of a neuron has a probabilistic interpretation when using sigmoid activations.

☐ A learning algorithm can scale the weights of a sigmoid-activation perceptron to approximate a threshold activation to arbitrary precision (as quantified by the maximum deviation from the threshold function).

☑ Gradient descent cannot be used on threshold activations.

Question 6 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 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.

○
Negative definite

○
Negative semidefinite

●
Positive definite

○
Indefinite

○
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 **(1.7, -1.2)** 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

$y =$ -1.5

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).

□ 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.

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

□ 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 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.

□ 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.

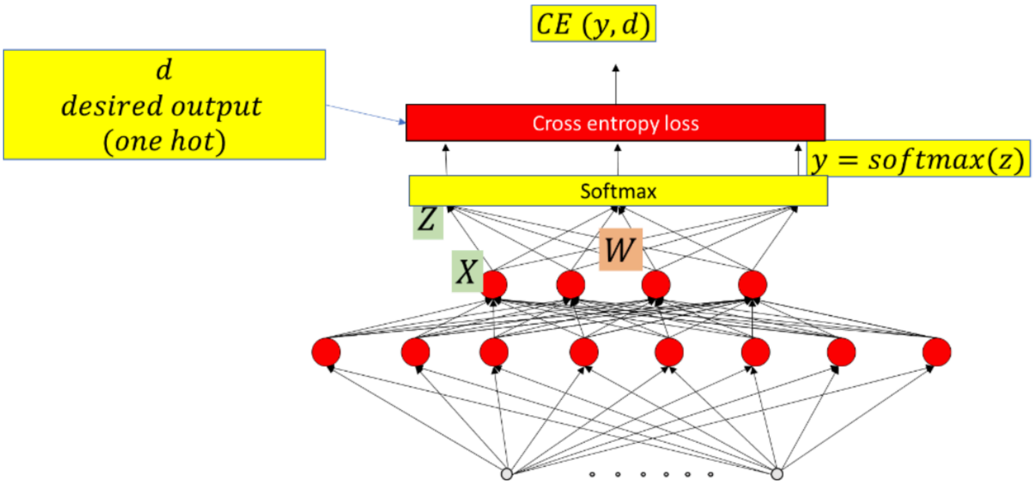
□ The derivative of a function $f(x)$ with respect to a variable z tells you how much minor perturbations of z perturbs $f(x)$

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.



0.73

No new data to save. Last checked at 2:06pm

Submit Quiz