Quiz 2

Started: Jan 25 at 9:31pm

Quiz Instructions

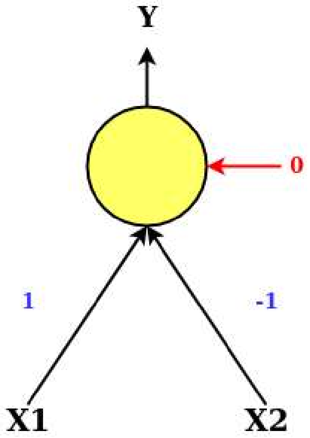
Learning in neural nets

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 perceptron:



X1 and **X2** are the inputs to the network. **Y** is the output of the network. The weights of the connections are shown in blue against the corresponding black arrows. The biases are shown in red. The perceptron uses the threshold activation function:

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

If the inputs to the perceptron are **X1**=1 and **X2**=1.25, the desired output is **d**=1and the weights and bias are updated using the ADALINE learning rule with a learning rate of η=0.4, then what will be the values of the weights and the bias after the 1st iteration? (Assume that **W1** is the weight associated with **X1** and **W2** is the weight associated with **X2**)

Hint: See hidden ADALINE and MADALINE slides. Lecture 3 slide 80-83

○
W1=1, **W2**=-1, bias=0

○
W1=1, **W2**=-1, bias=0.5

○ **W1**=0.5, **W2**=-1.625, bias=0.5

○ **W1**=1.5, **W2**=0.375, bias=-0.5

○ **W1**=1.5, **W2**=-0.375, bias=0.5

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 48-50

○ True

○ False

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

□ 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.

□ 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.

□ Threshold activations are inadequate to approximate most functions.

Question 4 1 pts

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

Hint: See slide Lec3 slide 80 - 92

□

MADALINE utilizes ADALINE to update neuron parameters

□

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

□

MADALINE is simply ADALINE, when it utilizes parallel computation

□

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 pts

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

Hint: See slide Lec3 p79 - 91

□

It ignores the threshold activation during training, and only applies it during testing.

□

It tries to minimize the error between the desired binary output and the affine combination of inputs before the threshold activation is applied.

□

It computes the squared error between the output of the perceptron and the target output, instead of counting errors.

□

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 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

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 $f(x_1, x_2, x_3) = x_1^2 x_2 + x_2^2 x_3 + x_3^3 + 2x_1 x_3 + x_2 x_3 + 6$ at the point (0,0,0) is :

Hint: See lec 4, slide 19, 34-37, and rewatch that portion of the lecture. You will have to work out the Hessian and compute its Eigenvalues.

○
Positive semidefinite

○
Negative semidefinite

○
Negative definite

○
Indefinite

○
Positive definite

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.

$x =$ □

$y =$ □

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).

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

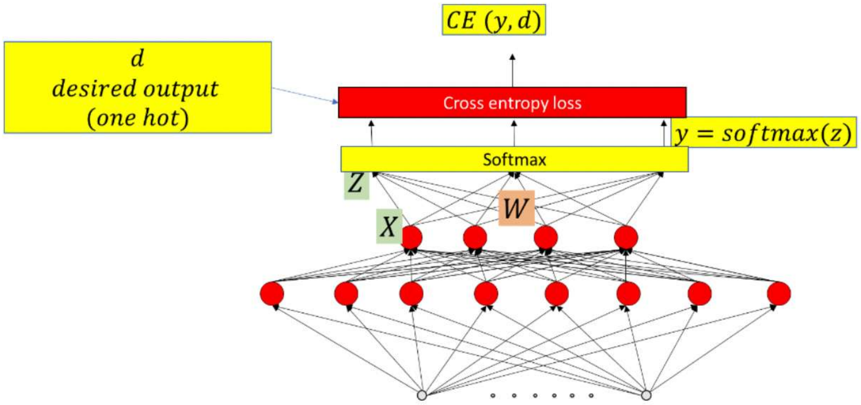
□
The derivative of a function $y = f(x)$ with respect to its input x is the ratio $\frac{dy}{dx}$ of small increments in the output that result from small increments of the input.

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

□
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.

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 below. This is followed by a weight matrix W which computes an affine value Z (also often called "logits") to which a softmax activation is applied to compute class probabilities.



What is the size of the weight matrix W. Using Python notation, assume all vectors are **row** vectors (i.e. X is a 1 x 4 vector). Note that this is different from the notation in class where all vectors are column vectors.

Hint: Lecture 4 slides 98-99

○ We cannot really say without additional specification

○ 4 x 3

○ 4 x 4

○ 3 x 4

Not saved Submit Quiz