Quiz-03

- Due Feb 2 at 11:59pm
- Points 10
- Questions 10
- Available Jan 31 at 6pm - Feb 2 at 11:59pm
- Time Limit None
- Allowed Attempts 3

Instructions

This quiz primarily covers lectures 5-6, but you are expected to be familiar with concepts from previous lectures as well.

Several of the questions refer to hidden slides that were not presented in class.

Some of the questions also require you to read additional material, links to which are posted in the quiz questions.

Take the Quiz Again

Attempt History

	Attempt	Time	Score
LATEST	Attempt 1	312 minutes	9 out of 10

Correct answers are hidden.

Score for this attempt: 9 out of 10
Submitted Feb 1 at 6:28pm
This attempt took 312 minutes.

Question 1
1 / 1 pts

For this question, please read the paper: **Rumelhart, Hinton and Williams (1986 (http://www.cs.toronto.edu/~hinton/absps/naturebp.pdf)) (http://www.cs.toronto.edu/~hinton/absps/naturebp.pdf)** .

[Can be found at: http://www.cs.toronto.edu/~hinton/absps/naturebp.pdf]

One version of gradient descent changes each weight by an amount proportional to the accumulated δE/ δw.

$$\Delta w = -\epsilon \frac{\delta E}{\delta w}$$

Select all that are true about this method:

☑ It can be improved without sacrificing simplicity and locality.

"It can be significantly improved, without sacrificing the simplicity and locality, [...]" p535

☑ It’s simpler than methods that use second derivatives.

"This method does not converge as rapidly as methods which make use of the second derivatives, but it is much simpler [...]" p535

☐ It cannot be implemented by local computations in parallel hardware.

☐ This method converges as rapidly as methods that make use of second derivatives.

Question 2

1 / 1 pts

(Select all that apply) As discussed in lecture, which of the following is true for the backpropagation algorithm?

Hint: Lecture 5, starting at "training by backprop".

☑ It cannot be performed without first doing a feed-forward pass of the input(s) through the network

☑ It is used to compute derivatives that are required for the gradient descent algorithm that trains the network

☑ It can be used to compute the derivative of the divergence with respect to the input of the network

☑ It computes the derivative of the divergence between the true and desired outputs of the network for a training input

☐ It computes the derivative of the average divergence for a batch of inputs

Question 3

1 / 1 pts

We are given a binary classification problem where the training data from both classes are linearly separable. We compare a perceptron, trained using the perceptron learning rule with a sigmoid-activation perceptron, trained using gradient descent that minimizes the L2 Loss. In both cases, we restrict the weights vector of the perceptron to have finite length. In all cases, we will say the algorithm has found a “correct” solution if the learned model is able to correctly classify the training data. Which of the following statements are true (select all that are true).

Hint: See slides 13-32, lecture 6

☑ There are situations where the gradient-descent algorithm will not find the correct solution.

☐ We cannot make any statement about the truth of falsity of the other options provided, based only on the information provided.

☑ The perceptron algorithm will always find the correct solution.

☐ The gradient-descent algorithm will always find the correct solution.

Question 4

1 / 1 pts

(Select all that apply) At any point, the gradient of a scalar function with multivariate inputs…

Hint: Lecture 4, "Gradient of a scalar function of a vector" and "properties of a gradient".

☑ Is in the direction of steepest ascent

☐ Is in the direction of steepest descent

☑ Is the vector of local partial derivatives w.r.t. all the inputs

☐ Is parallel to equal-value contours of the function

Gradient points in the direction of steepest ascent, hence you take a step in the negative gradient for gradient descent “parallel to equal-value contours” refers to the direction where loss is the same (wrong)

“Is the vector of local partial derivatives w.r.t. all the inputs”: “inputs” in the context of neural nets would refer to the weights and biases, not the actual input of the neural net

Question 5

1 / 1 pts

Let *d* be a scalar-valued function with multivariate input, *f* be a vector-valued function with multivariate input, and *X* be a vector such that y = d(f(X)). Using the lecture’s notation, assuming the output of *f* to be a column vector, the derivative $\nabla_f y$ of y with respect to f(X) is…

Hint: (Lecture 4 and) Lecture 5, Vector calculus, Notes 1.

☐ Composed of the partial derivatives of y w.r.t the components of X

☐ A column vector

☑ A row vector

☐ A matrix

Question 6

1 / 1 pts

Which of the following are valid subgradients of a RELU, given by Relu(x), at x = 0? We will represent the subgradient as $\nabla_{subgrad} RELU\left(x\right)$

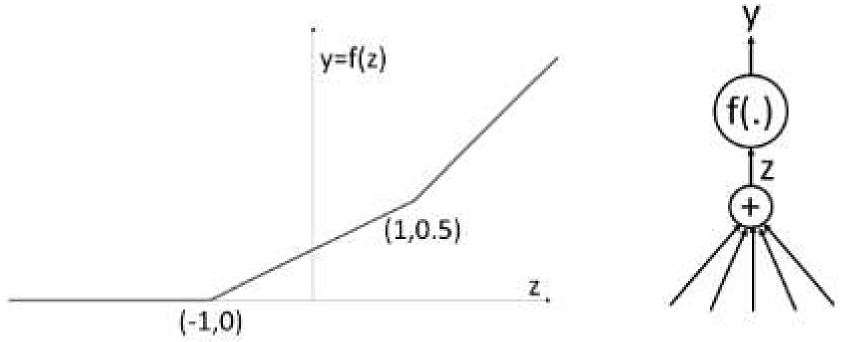
Hint: Lecture 5, slides 112-114.

Hint: When a function is not differentiable at a certain point, we use subgradients to represent valid slopes or directions at that point.

☑ $\nabla_{subgrad} RELU\left(0\right) = 1$
☑ $\nabla_{subgrad} RELU\left(0\right) = 0$
☑ $\nabla_{subgrad} RELU\left(0\right) = 0.5$
☐ $\nabla_{subgrad} RELU\left(0\right) = -0.5$
☐ $\nabla_{subgrad} RELU\left(0\right) = 1.5$

IncorrectQuestion 7
0 / 1 pts

The following piecewise linear function with “hinges” at (-1,0) and (1,0.5) is used as an activation for a neuron. The slope of the last segment is 40 degrees with respect to the *z* axis (going anti-clockwise). Our objective is to find a z that minimizes the divergence div(y,d). Which of the following update rules is a valid subgradient descent update rule at z=1? Here η is the step size and is a positive number. The superscript on *z* represents the step index in an iterative estimate. The derivative $\frac{\partial div(y,\, d)}{\partial\ z}$ is computed at $z^k = 1$. The value of η must not factor into your answer (i.e. remember that η has only been included in the equations for completeness sake and do not argue with us that you can always adjust η to make any answer correct ☺)



Hint: Lecture 5, slides 112-114

☑ $z^{k+1} = z^k - \eta 0.1 \frac{\partial div(y,d)}{\partial y}$
☐ $z^{k+1} = z^k + \eta \frac{\partial div(y,d)}{\partial y}$
☑ $z^{k+1} = z^k - \eta \frac{\partial div(y,d)}{\partial y}$
☑ $z^{k+1} = z^k - \eta 0.75 \frac{\partial div(y,d)}{\partial y}$
☑ $z^{k+1} = z^k - \eta 0.25 \frac{\partial div(y,d)}{\partial y}$

Check Lecture 5, slides 112-114: The correct choices for n are those that align with the slope of the piecewise linear function at z = 1

Question 8

1 / 1 pts

In order to maximize the possibility of escaping local minima and finding the global minimum of a generic function, the best strategy to manage step sizes during gradient descent is:

Hint: Lecture 6, "Issues 2"

○ To maintain a step size consistently close to the optimal step size (e.g. close to the inverse second derivative at the current estimate)

○ To keep the step size low throughout to prevent divergence into a local minima

● To start with a large, divergent step size (e.g. greater than twice the optimal step size for a quadratic approximation at the initial location) and gradually decrease it over iterations

○ To start with a large, non-divergent step size (e.g. less than twice the optimal step size for a quadratic approximation at the initial location) and gradually decrease it over iterations

See lecture for explanation.

Question 9

1 / 1 pts

Which of the following update rules explicitly computes second-order derivatives or their approximations? (select all that apply)

Hint: second half of Lecture 6

☑ Newton's Method

□ RProp

□ Gradient descent

☑ Quickprop

Question 10

1 / 1 pts

Let f be a quadratic function such that at $x = 1$, $f(x) = 10$, $f'(x) = -4$, and $f''(x) = 1$. The minimum has a value of x = 5.0 and a value of $f(x)$ = 2.0 . (Truncate your answer to 1 digit after the decimal point i.e. enter your answer in the format x.x, e.g. 4.5)

Hint: Lecture 6 "Convergence for quadratic surfaces"

Answer 1:

5.0

Answer 2:

2.0

- Need to change answer on canvas to accept decimals

Quiz Score: 9 out of 10