Quiz-04

Quiz submitted

- Due Feb 9 at 11:59pm
- Points 10
- Questions 10
- Available Feb 7 at 6pm - Feb 9 at 11:59pm
- Time Limit None
- Allowed Attempts 3

Take the Quiz Again

Attempt History

	Attempt	Time	Score
LATEST	Attempt 1	225 minutes	7 out of 10

Correct answers are hidden.

Score for this attempt: 7 out of 10
Submitted Feb 8 at 11:14pm
This attempt took 225 minutes.

Incorrect Question 1
0 / 1 pts

You are trying to minimize the cross-entropy loss of the logistic function $y = \frac{1}{1+exp(0.5w)}$ with respect to parameter w, when the target output is 1.0. Note that this corresponds to optimizing the logistic function $y = \frac{1}{1+exp(-wx)}$, given only the training input (x, d(x)) = (-0.5, 1). You are using Nestorov's updates. Your current estimate (in the k-th step) is $w^{(k)} = 0$. The last step you took was $\Delta w^{(k)} = 0.5$. Using the notation from class, you use $\beta = 0.9$ and $\eta = 0.1$. What is the value of $w^{(k+1)}$ when using Nestorov's update? Truncate the answer to three decimals (**do not round up**).

Hint: The cross entropy loss is identical to the KL divergence (lec8, "choices for divergence"), when the target output is binary (or, more generally, one hot).

0.425

Question 2
1 / 1 pts

Quiz submitted

Several researchers separately decide to estimate an unknown function d(x), for a variable x. Although they do not know the function, they do have access to an oracle who does know d(x). Upon demand the oracle will randomly draw a value x from a ***uniform*** probability distribution $P(x) = 1, 0 \leq x \leq 1$ and return (x, d(x)), i.e. a random value of x and the value of the function d(x) at that x. Each of the researchers independently obtains 1000 training pairs from the oracle, and begins to estimate d(x) from them. They begin with an initial estimate of d(x) as y(x)=0 (where y(x) is the estimate of d(x). They do not update y(x) during this exercise). Then each of them computes the average L2 divergence (as defined in lecture 4) over their entire batch of training examples.

In order to get a better handle on their problem, they pool their divergences. Since each of them got an independent set of training samples, all their divergences are different. So they compute the statistics of the collection of divergences.

Unknown to them (but known to the oracle), $d(x) = \sqrt{x}$

What would you expect the average of their pooled set of divergences to be? Truncate the answer to 2 decimals (and write both digits, even if the answer has the form *.00)

Hint: Lec 7, slides 50-80. The L2 divergence is as defined in lec 8, “choices for divergence”, i.e ½ times the squared Euclidean error.

0.25

Question 3
1 / 1 pts

Several researchers separately decide to estimate an unknown function d(x), for a variable x. Although they do not know the function, they do have access to an oracle who does know d(x). Upon demand the oracle will randomly draw a value x from a ***Exponential*** probability distribution $P(x) = Exponential(\lambda = 0.1)$ and return (x,d(x)), i.e. a random value of x and the value of the function d(x) at that x. Each of the researchers independently obtains 1000 training pairs from the oracle, and begins to estimate d(x) from them. They begin with an initial estimate of d(x) as y(x)=0 (where y(x) is the estimate of d(x). They do not update y(x) during this exercise)). They plan to process their training data using stochastic gradient descent. So each of them computes the L2 divergence (as defined in lecture 4) between the estimated output and the desired output for each of their training instances. They do not update y(x) during this exercise.

In order to get a better handle on their problem, the researchers then get together and pool their divergences over all of their combined training instances, and compute the statistics of the collection of divergences.

Unknown to them (but known to the oracle) $d(x) = \sqrt{x}$

What would you

Quiz submitted

Note: Recall that a random variable $X \sim Exponential(\lambda)$ if the pdf is $p_X(x) = \lambda exp(-\lambda x)$ for x belongs to [0,infinity). The expectation of the random variable is given by E[X] =1/λ and the variance of the random variable is given by Var(X) = 1/λ^2.

Truncate the answer to 1 decimals (and write the number after the decimal, even if the answer has the form *.0)

Extra Hint: Lec 7, slide 50-80. The L2 divergence is as defined in lec 8, “choices for divergence”, i.e ½ times the squared Euclidean error.

25.0

Please check Lecture 7

Question 4
1 / 1 pts

What could happen if incremental gradient descent was not stochastic (i.e. if we selected the training points in a constant order) ?

- A function that swings around instead of converging
- The loss value would converge very quickly
- The behavior would not change
- Overfitting

Explanation: See lec 7 slides on order of presentation

See lec 7 slides on order of presentation

IncorrectQuestion 5
0 / 1 pts

Compared to batch gradient descent, SGD is often faster (takes less time) because:

- it needs about the same number of iterations, but they are faster to compute
- its iterations take longer to compute, but it converges in fewer iterations
- we get many more updates in a single pass through the training data
- it is not faster

See lec 7 slides on SGD convergence

Quiz submitted

Question 6

1 / 1 pts

(select all that apply) Which of the following is true of the optimal schedule of learning rates for incremental update methods?

Hint: Lec 7

- [ ] The sum of the squares of the learning rates must be infinite in the limit (of infinite iterations).
- [x] The sum of the learning rates must be infinite in the limit (of infinite iterations).
- [x] The learning rates must initially be large, and shrink with iterations
- [ ] The learning rates must always be less than twice the inverse of the second derivative to ensure convergence
- [ ] The sum of the learning rates must be finite in the limit (of infinite iterations), in order to ensure convergence
- [x] The sum of the squares of the learning rates must be finite in the limit (of infinite iterations)

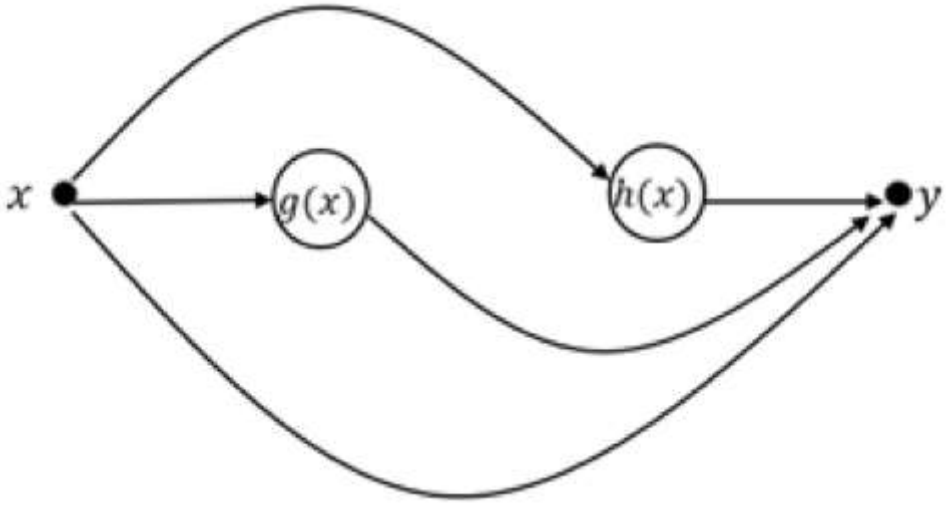
Question 7

1 / 1 pts

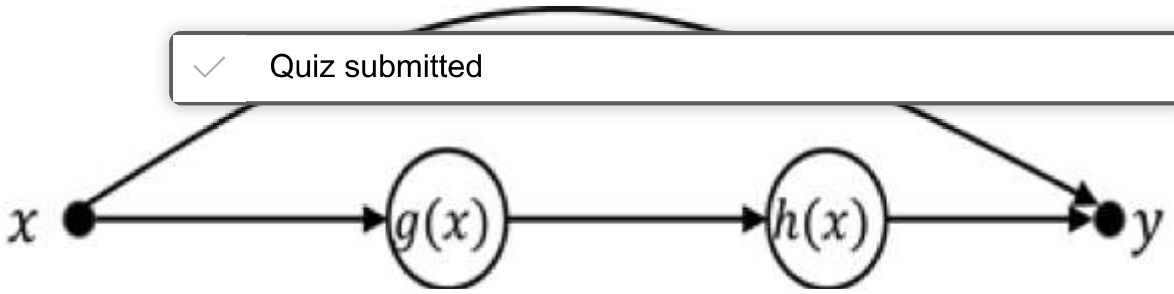
We are given the following relationship $y = f(x, g(x), h(x, g(x)))$. Which of the following figures is the influence diagram for *y* as a function of x.

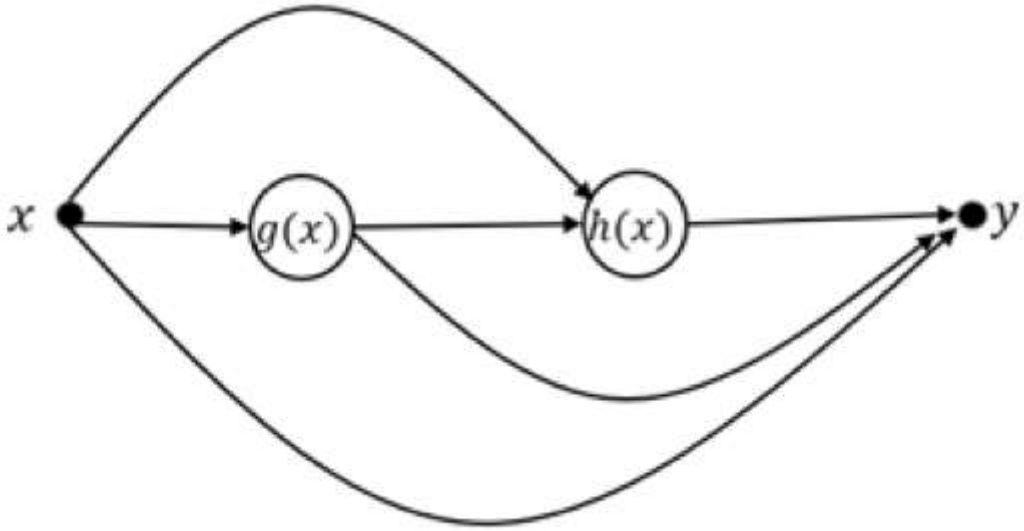
(In the figures below we have generically depicted the dependence between h() and x as h(x) as a shorthand notation.)

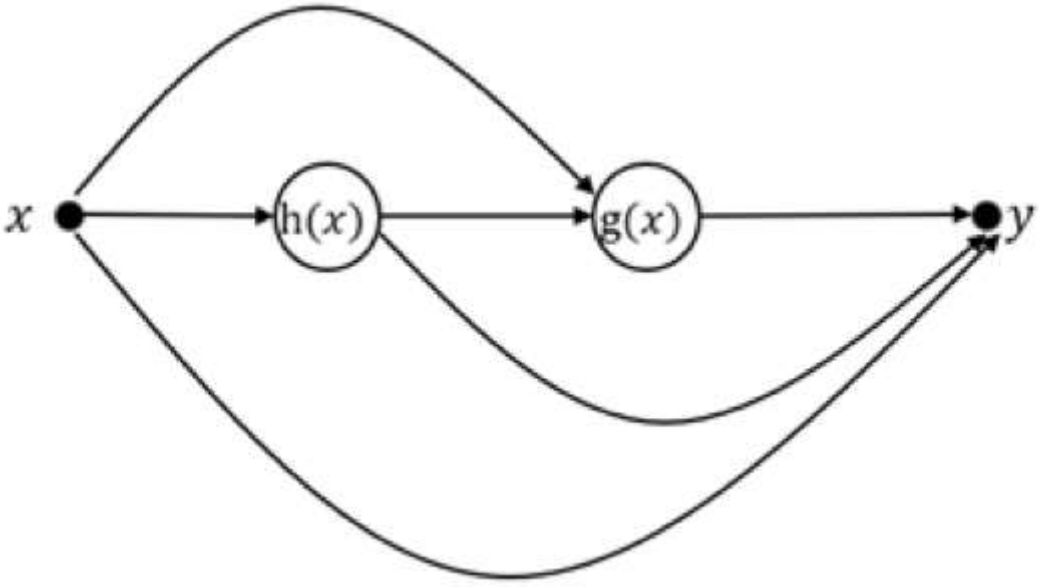
Hint: Lecture 5, Slides 11-23



Quiz submitted







PartialQuestion 8

0.5 / 1 pts

We are given the relationship $y = f(x, g(x), h(x, g(x)))$. Which of the following correctly specify the formula for the derivative of *y* w.r.t. x? You may find it useful to draw the influence diagram. Note that

partial derivatives are denoted using and full derivatives are denoted using d. **(Select all that apply)**

Quiz submitted

Hint: Lecture 5, Slides 11-32. Also note the difference between full and partial derivatives

- [ ] $dy/dx = \partial y/\partial x \times \partial h/\partial g \times dg/dx$
- [ ] $dy/dx = \partial y/\partial x + \partial y/\partial g \times \partial g/\partial x + \partial y/\partial h \times \partial h/\partial x$
- [ ] $dy/dx = \partial y/\partial x + \partial y/\partial g \times dg/dx + \partial y/\partial h \times dh/dx$
- [x] $dy/dx = \partial y/\partial x + \partial y/\partial g \times dg/dx + \partial y/\partial h \times (\partial h/\partial x + \partial h/\partial g \times dg/dx)$
- [ ] $dy/dx = \partial y/\partial x + \partial y/\partial g + \partial y/\partial h$

Question 9
1 / 1 pts

Dropout is a regularization technique, which theoretically emulates bagging, applied to a network with N (non-output) neurons. When we use Dropout, each (non-output) neuron in the network is selected (turned on) with a probability α. During inference, each of the neuron's activations is scaled by α to account for the dropout applied during training. Which of the following statements is true about this procedure?

Hint: Lec 8, Dropout slides, slide 132

- [ ] It utilizes the fact that that

$E[f(D_1 g_1(x), D_2 g_2(x), \ldots, D_N g_N(x))] = f(E[D_1 g_1(x)], E[D_2 g_2(x)], \ldots, E[D_N g_N(x)])$

where $f()$ represents the network, $g_i(x)$ is the i-th neuron in the network, and D_i is the Bernoulli selector for the neuron, and $E[.]$ is the expectation operator. This is a provably correct equality.

- [ ] It computes the output as

$y = f(E[D_1 g_1(x)], E[D_2 g_2(x)], \ldots, E[D_N g_N(x)])$

where $f()$ represents the network, $g_i(x)$ is the i-th neuron in the network, and D_i is the Bernoulli selector for the neuron, and $E[.]$ is the expectation operator. This is a theoretically precise computation of the output for bagging.

- [x] It makes the approximation that

$E[f(D_1 g_1(x), D_2 g_2(x), \ldots, D_N g_N(x))] \approx f(E[D_1 g_1(x)], E[D_2 g_2(x)], \ldots, E[D_N g_N(x)])$

where $f()$ represents the network, $g_i(x)$ is the i-th neuron in the network, and D_i is the Bernoulli selector for the neuron, and $E[.]$ is the expectation operator. This is only an approximation.

PartialQuestion 10

0.5 / 1 pts

Quiz submitted

To answer this question, please read (https://arxiv.org/abs/1502.03167 (https://arxiv.org/abs/1502.03167)).

As referred in the paper, in the BN transformation, (indicate all true options)

- [x] normalized inputs accelerate training of the network
- [x] it can manipulate only differentiable activation functions
- [x] gradients of the loss w.r.t inputs are computed by computing the gradients w.r.t the outputs of the BN transform as well as the gradients of the outputs of the BN transform w.r.t the pre-normalized inputs

Quiz Score: 7 out of 10