50.021 Artificial Intelligence

Quiz 2

Student Name: **Student ID:**

[Q1]. True or False (Give one sentence explanation) **[6p]**

a Gradient descent is a way to minimize an objective function $J(\theta)$ parameterized by a model's parameters $\theta \in \mathbb{R}^d$ by updating the parameters in the same direction of the gradient of the objective function $\nabla_\theta J(\theta)$ w.r.t. to the parameters.
Solution: False. We need to update the parameters in the opposite direction of the gradient objective function.

b Although gradient descent is one of the most popular algorithms to perform optimization, it can only find a local minimum which can be good or bad.

Solution: True. The only way to find a global minimum or maximum is to travel the entire hypothesis space which may take a very very long time.

c Choosing a proper learning rate in gradient descent can be difficult. A learning rate that is too small can hinder convergence and cause the loss function to fluctuate around the minimum.

Solution: False. A learning rate that is too small in a flat region can hinder convergence, but learning rate that is too large causes function to fluctuate around the minimum.

d The momentum term increases for dimensions whose gradients point in the same directions and reduces updates for dimensions whose gradients change directions. As a result, we gain faster convergence and reduced oscillation.
Solution: True. As the name suggests, the momentum term reduces speed for dimensions whose gradient change directions and vice versa.

e The XOR operator can be modeled using a one layer NN with a tanh/sigmoid nonlinearity.

Solution: False. We need at least two layers to model the XOR operator.

f In a neural network, knowing the weight and bias of each neuron is the most important step. If you can somehow get the correct value of weight and bias for each neuron, you can approximate any continuous function.
Solution: True. The training is to find the true value of weight and bias, that is assumed to exist to produce that specific continuous function.

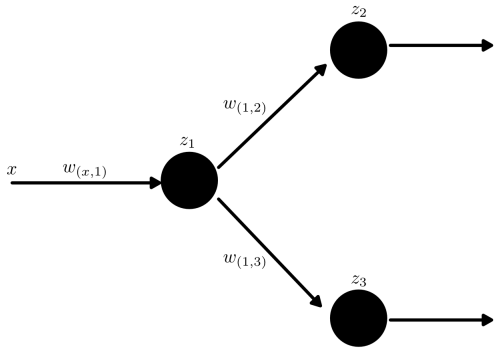

Figure 1: Mini Neural Network

[Q2]. Refer to the neural network at figure 1 with input $x \in \mathbb{R}^1$. The activation function for z_1, z_2, and z_3 is the sigmoid function: $\frac{1}{1+e^{-w \cdot x}}$,

$$h(x) = \frac{1}{1+e^{-x}} \tag{1}$$
$$z_1 = h(x \cdot w_{(x,1)}) \tag{2}$$
$$z_2 = h(z_1 \cdot w_{(1,2)}) \tag{3}$$
$$z_3 = h(z_1 \cdot w_{(1,3)}) \tag{4}$$

For the error E, instead of using the softmax function we learned in class, we use the quadratic error function for regression purpose,

$$E = \sum_{i \in data} ((z_2 - y_{2i})^2 + (z_3 - y_{3i})^2)$$

[6p] Write down an expression for the gradients of all three weights: $\frac{\partial E}{\partial w_{(x,1)}}, \frac{\partial E}{\partial w_{(1,2)}}, \frac{\partial E}{\partial w_{(1,3)}}$.

Solution:

We have the following equations from the NN:

$$\frac{\partial E}{\partial z_j} = \sum_{i \in data} 2(z_j - y_{ij}) \tag{5}$$
$$\frac{\partial z_i}{z_j} = (1 - h(z_j \cdot w_{(j,i)})) \cdot w_{(j,i)} \tag{6}$$
$$\frac{\partial z_i}{w_{(j,i)}} = (1 - h(z_j \cdot w_{(j,i)})) \cdot z_j \tag{7}$$
$$\tag{8}$$

Then the update rules are,

$$\frac{\partial E}{\partial w_{(1,3)}} = \frac{\partial E}{\partial z_3} \frac{\partial z_3}{\partial w_{(1,3)}}$$
$$\frac{\partial E}{\partial w_{(1,2)}} = \frac{\partial E}{\partial z_2} \frac{\partial z_2}{\partial w_{(1,2)}}$$
$$\frac{\partial E}{\partial w_{(x,1)}} = \frac{\partial E}{\partial z_1} \frac{\partial z_1}{\partial w_{(x,1)}}$$
$$= (\frac{\partial E}{\partial z_2} \frac{\partial z_2}{\partial z_1} + \frac{\partial E}{\partial z_3} \frac{\partial z_3}{\partial z_1}) \frac{\partial z_1}{\partial w_{(x,1)}}$$

Each term can be found from equations (1) to (8) above.