Quiz 5

● Graded

Student

HARRIS DOAN

Total Points

18 / 18 pts

Question 1

(no title) **2** / 2 pts

✔ **+ 2 pts** Correct

Question 2

(no title) **1** / 1 pt

✔ **+ 1 pt** Correct

Question 3

(no title) **1** / 1 pt

✔ **+ 1 pt** Correct

Question 4

(no title) **2** / 2 pts

✔ **+ 2 pts** Correct

Question 5

(no title) **2** / 2 pts

✔ **+ 2 pts** Correct

Question 6

(no title) **1** / 1 pt

✔ **+ 1 pt** Correct

Question 7

(no title) **2** / 2 pts

✔ **+ 2 pts** Correct

Question 8

(no title) **1** / 1 pt

✔ **+ 1 pt** Correct

Question 9

(no title) **2** / 2 pts

✔ **+ 2 pts** Correct

Question 10

(no title) **2** / 2 pts

✔ **+ 2 pts** Correct

Question 11

(no title) **2** / 2 pts

✔ **+ 2 pts** Correct

Q1
2 Points

Variables $a, b, c, d, e, f \in \mathbb{R}$ satisfy

$$c = \text{ReLU}(w_1 \cdot a + w_2 \cdot b)$$
$$d = \tanh(w_3 \cdot a + w_4 \cdot b)$$
$$e = \sigma(w_5 \cdot a + w_6 \cdot b)$$
$$f = \text{ReLU}(w_7 \cdot d + w_8 \cdot e)$$
$$g = \text{ReLU}(w_9 \cdot c + w_0 \cdot f)$$

where w_i is a constant for all $i \in \{0, \ldots, 9\}$.

Which one of the following statements is true?

Hint: It may be helpful to draw a computational graph.

- (●) $\frac{\partial g}{\partial b} = \frac{\partial g}{\partial c} \cdot \frac{\partial c}{\partial b} + \frac{\partial g}{\partial f} \cdot \frac{\partial f}{\partial b}$
- ( ) $\frac{\partial g}{\partial b} = \frac{\partial g}{\partial c} \cdot \frac{\partial c}{\partial b} + \frac{\partial g}{\partial f} \cdot \frac{\partial f}{\partial d} \cdot \frac{\partial d}{\partial b}$
- ( ) None of the above
- ( ) $\frac{\partial g}{\partial a} = \frac{\partial g}{\partial c} \cdot \frac{\partial c}{\partial a}$
- ( ) $\frac{\partial f}{\partial a} = \frac{\partial d}{\partial a} + \frac{\partial e}{\partial a}$

Q2
1 Point

While a single-layer perceptron cannot perfectly classify a dataset that is not linearly separable, there exists a 5-layer neural net with a linear activation function at every unit that can perfectly classify such a dataset.

- ( ) True
- (●) False

Q3

1 Point

We are attempting to train a neural network with a single hidden layer using gradient descent. We use sigmoid for all the activation functions in the hidden layer. The learned parameter values will **never** depend on their initialization.

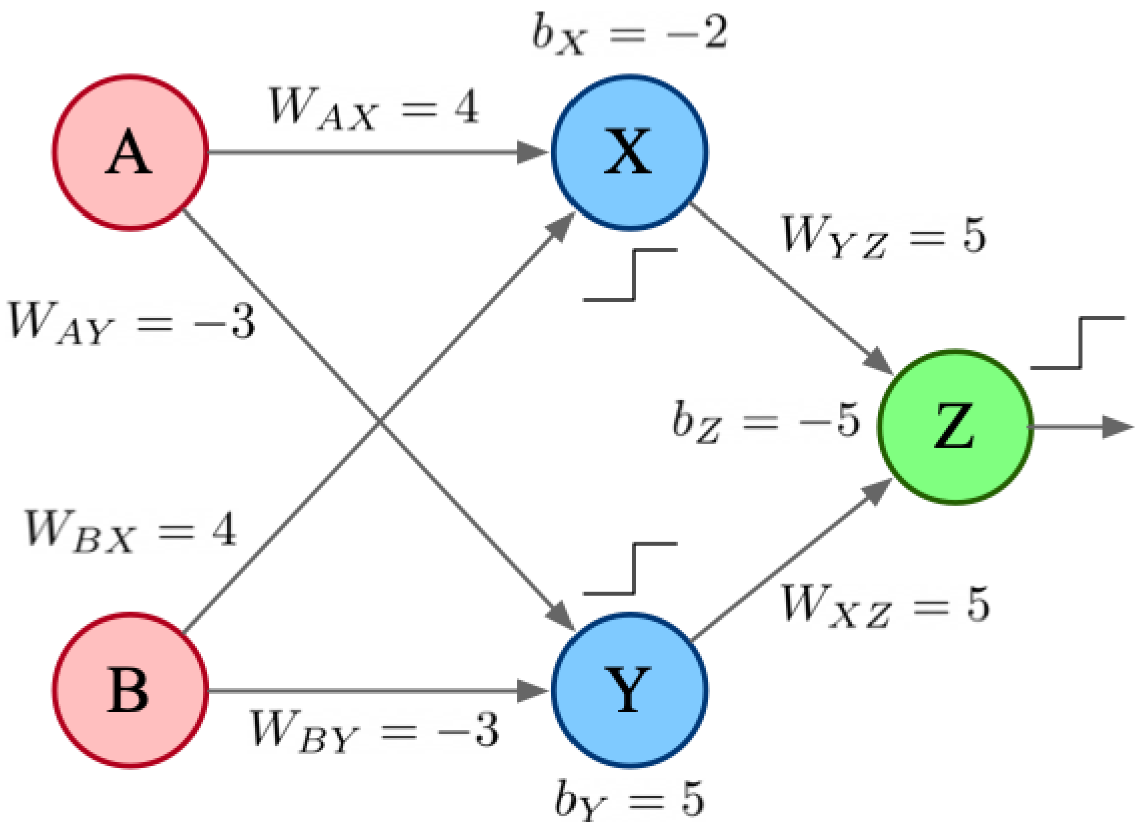
- (●) False
- ( ) True

Q4

2 Points

Consider the following neural network, which contains two input units (A and B), two hidden units (X and Y), and one output unit (Z). The value associated with each weight w_{ij} and bias b_i term is given in the figure below. Note that the activation function used here is the step function, defined as

$$\text{step}(x) = \begin{cases} 1, & x > 0 \\ 0, & x \leq 0 \end{cases}$$



Assuming the inputs only take binary values (i.e. $\{0, 1\}$), which logical operation does this neural network represent?

Note: For truth tables of logical operations, please check out https://en.wikipedia.org/wiki/Logic_gate

- AND
- None of above
- NOR
- OR
- (selected) XOR

Q5
2 Points

A neural network has one input layer $\mathbf{x}$ with 5 neurons, one hidden layer $\mathbf{h} = f(\mathbf{W_1 x} + \mathbf{b_1})$ with 10 neurons, and one output layer $\mathbf{z} = f(\mathbf{W_2 h} + \mathbf{b_2})$ with 3 neurons.

What is the total number of learnable parameters (i.e. weights and biases)?
Only enter an integer number (e.g. 157) for your answer.

93

Q6
1 Point

We are using gradient descent to learn the parameters of a simple neural network for binary classification: $f(x) = \sigma(w_1 x + w_0)$, where $x, w_0, w_1 \in \mathbb{R}$ and σ is the sigmoid function.

We are **more likely** to encounter the problem of vanishing gradients if we initialize the parameters (w_0, w_1) to large values.

- False
- (selected) True

Q7

2 Points

Which one of the following statements about kernel functions is **incorrect**?

- ○ For any kernel function $k(\mathbf{u}, \mathbf{v})$, $k(\mathbf{u}, \mathbf{v}) = \phi(\mathbf{u})^T \phi(\mathbf{v})$ for some function ϕ.
- ○ Some kernel functions have their own hyperparameters to tune.
- ◉ Any bivariate function $f(\mathbf{u}, \mathbf{v})$ is a valid kernel function.
- ○ Using kernel functions can make nonlinear classifiers more computationally efficient.

Q8

1 Point

Given any two kernel functions $k_1(\mathbf{u}, \mathbf{v})$ and $k_2(\mathbf{u}, \mathbf{v})$ that take vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^2$ as input, $7k_1(\mathbf{u}, \mathbf{v}) + 3k_2(\mathbf{u}, \mathbf{v}) - 1$ is **always** a valid kernel function.

- ○ True
- ◉ False

Q9

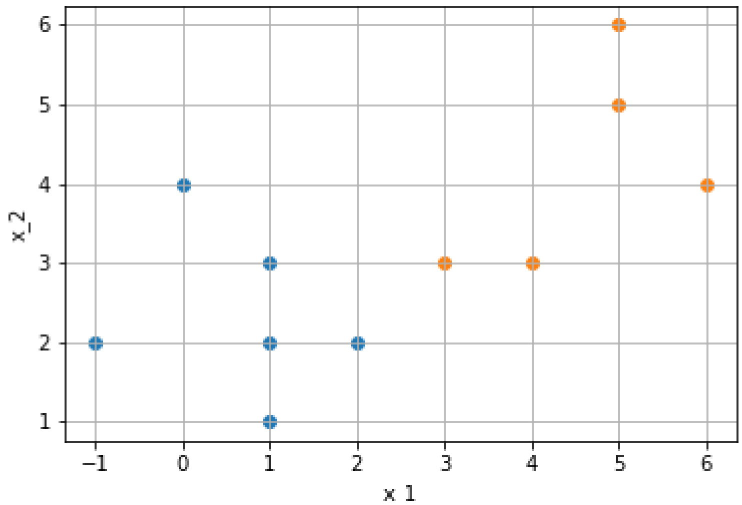
2 Points

Assume that the dataset is not linearly separable, and SVM with slack variables still makes a few misclassifications after optimization. The value of the slack variable ξ_1 for a misclassified point x_1 would satisfy:

- ○ $\xi_1 = 0$
- ◉ $\xi_1 \geq 1$
- ○ $0 < \xi_1 < 1$
- ○ $\xi_1 < 0$

Q10
2 Points

Suppose we have collected the 2D samples plotted below,



What would be the boundary computed by the hard-margin SVM?

- ◯ $x_1 - 2.5 = 0$
- ◉ $5 - x_1 - x_2 = 0$
- ◯ $x_2 + 2.5 = 0$
- ◯ None of the above
- ◯ $x_1 - x_2 - 2 = 0$

Q11
2 Points

We can introduce non-linearity to SVM using the kernel trick. Instead of searching for a hyperplane $\mathbf{w}^T\mathbf{x} + b$ that maximizes the margin, we are looking for $\mathbf{w}^T\phi(\mathbf{x}) + b$, where ϕ is the non-linear basis function. We are given training data $\{(\mathbf{x}_n, y_n)\}$ to learn the kernel SVM.

Which one of the following statements is **wrong** about kernel SVM?

- ○ We can predict the label of a new sample using the kernel function and the training data.
- ○ If we apply an appropriate kernel function, non-separable data **may** become separable.
- ◉ The suport vectors are the instances where the dual variable (α) is zero.
- ○ A valid kernel function should have a positive-semidefinite kernel matrix.