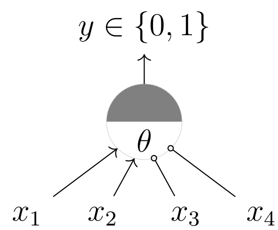
1. Recall that McCulloch Pitts (MP) neuron aggregates the inputs and takes a decision based on this aggregation. If the sum of all inputs is greater than the threshold (θ), then the output of MP neuron is 1, otherwise the output is 0. We say that a MP neuron implements a boolean function if the output of the MP neuron is consistent with the truth table of the boolean function. In other words, if for a given input configuration, the boolean function outputs 1 then the output of the neuron should also be 1. Similarly, if for a given input configuration, the boolean function outputs 0 then the output of the neuron should also be 0.
   Consider the following boolean function:

$$f(x_1, x_2, x_3, x_4) = (x_1 \text{ AND } x_2) \text{ AND } (!x_3 \text{ AND } !x_4)$$

   The MP neuron for the above boolean function is as follows:



   What should be the value of the threshold (θ) such that the MP neuron implements the above boolean function? (Note that the circle at the end of the input to the MP neuron indicates inhibitory input. If any inhibitory input is 1 the output will be 0.)

   A. $\theta = 1$
   B. $\theta = 2$
   C. $\theta = 3$
   D. $\theta = 4$

Solution: MP neuron will output 1 if and only if the sum of it's input is greater than or equal to the threshold, i.e.

$$\sum_{i=1}^{4} x_i \geq \theta$$

There are 16 possible inputs to this network: {0,0,0,0}, {0,0,0,1}, ..., {1,1,1,1}. Since x_3, x_4 are are inhibitory inputs, the output of the neuron will be zero whenever any one of these inputs is 1. This is as expected, because the output of the boolean function will also be 0 when either x_3 or x_4 is 1. There are 8 such inputs where either x_3 or x_4 will be 1 and hence the output of the neuron will be 0. Out of the remaining 8 inputs, the given boolean function $f(x_1, x_2, x_3, x_4)$ will output 1 if and only if $x_1 = 1, x_2 = 1, x_3 = 0, x_4 = 0$ and 0 for all other cases. Thus, we want the MP neuron to fire only if both x_1 and x_2 are 1, *i.e.*, we want the sum

$$x_1 + x_2 + x_3 + x_4 \geq 2$$

which implies

$$\theta = 2$$

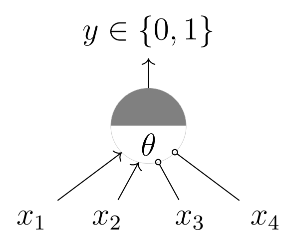
.

$\therefore$ **Option B** is the correct answer.

2. Keeping the concept discussed in question 1 in mind, consider the following boolean function:

$$f(x_1, x_2, x_3, x_4) = (x_1 \text{ OR } x_2) \text{ AND } (!x_3 \text{ AND } !x_4)$$

The MP neuron for the above boolean function is as follows:



What should be the value of the threshold (θ) such that the MP neuron implements the above boolean function? (Note that the circle at the end of the input to the MP neuron indicates inhibitory input. If any inhibitory input is 1 the output will be 0.)

A. $\theta = 1$

B. $\theta = 2$

C. $\theta = 3$

D. $\theta = 4$

Solution: MP neuron will output 1 if and only if the sum of it's input is greater than or equal to the threshold, i.e.

$$\sum_{i=1}^{4} x_i \geq \theta$$

There are 16 possible inputs to this network: {0,0,0,0}, {0,0,0,1}, ..., {1,1,1,1}. Since x_3, x_4 are are inhibitory inputs, the output of the neuron will be zero whenever any one of these inputs is 1. This is as expected, because the output of the boolean function will also be 0 when either x_3 or x_4 is 1. There are 8 such inputs where either x_3 or x_4 will be 1 and hence the output of the neuron will be 0. Out of the remaining 8 inputs, the given boolean function $f(x_1, x_2, x_3, x_4)$ will output 1 for any one of the following input settings:

$$x_1 = 1, x_2 = 0, x_3 = 0, x_4 = 0$$

$$OR$$

$$x_1 = 0, x_2 = 1, x_3 = 0, x_4 = 0$$

and 0 for all other input settings. Thus, we want the MP neuron to fire if either x_1 or x_2 is 1, *i.e.*, we want the sum

$$x_1 + x_2 + x_3 + x_4 \geq 1$$

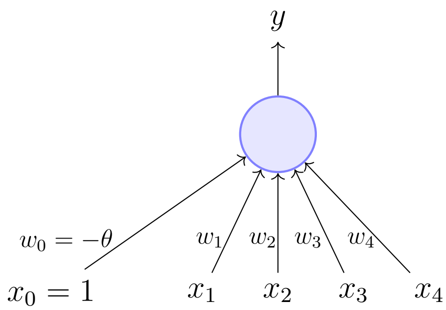
which implies

$$\theta = 1$$

.

$\therefore$ **Option A** is the correct answer.

3. Let us consider the movie example as discussed in this week's lecture. Suppose we want to predict whether a movie buff would like to watch a movie or not. Note that each movie is represented by a vector, $\mathbf{X} = [x_1 \ x_2 \ x_3 \ x_4]$ and the description of each input (x_i) is mentioned in the figure below. Also, the weight assigned to each of these inputs (or features) is given by $\mathbf{W} = [w_1 \ w_2 \ w_3 \ w_4]$ and the threshold is represented by the parameter θ.



$$x_1 = \textit{popularity(between 1 to 10)}$$
$$x_2 = \textit{isGenreScifi(boolean)}$$
$$x_3 = \textit{isDirectorNolan(boolean)}$$
$$x_4 = \textit{imdbRating(between 0 to 1)}$$

Now, consider the movie **Interstellar** has the feature vector $\mathbf{X} = [8 \;\; 1 \;\; 1 \;\; 0.86]$; which means the movie has a *popularity* of 8 on a a scale of 10 and is a *Scifi* movie directed by *Nolan* with 0.86 as its *imdbRating*. Now consider a person who assigns the following weights to each of these inputs: $\mathbf{W} = [0.14 \;\; 1 \;\; 0.9 \;\; 0.6]$. Further, suppose that $\theta = 2$. Based on the above information, what do you think will be his/her decision?

A. Yes, (s)he will watch it.

B. No, (s)he won't watch it.

Solution: His/Her decision will be to watch the movie if the weighted sum of the inputs is greater than the threshold, ie.

$$\sum_{i=1}^{4} w_i x_i \geq \theta$$

In this case, we have,

$$\begin{aligned}\sum_{i=1}^{4} w_i x_i &= (8 \times 0.14) + (1 \times 1) + (1 \times 0.9) + (0.86 \times 0.6) \\ &= 3.536 \\ &\geq \theta = 2\end{aligned}$$

which means (s)he will watch the movie.
$\therefore$ **Option A** is the correct answer.

4. Keeping the discussion of question 3 in mind, consider the movie **The Green Lantern** has the feature vector $\mathbf{X} = [5 \;\; 1 \;\; 0 \;\; 0.53]$. Now consider a person who assigns the following weights to each of these inputs: $\mathbf{W} = [0.8 \;\; 1 \;\; 0.4 \;\; 0.8]$. Further, suppose that $\theta = 7$.
Based on the above information, what do you think will be his/her decision?

A. Yes, (s)he will watch it.

B. No, (s)he won't watch it.

Solution: In this case,

$$\sum_{i=1}^{4} w_i x_i = (5 \times 0.8) + (1 \times 1) + (0 \times 0.4) + (0.53 \times 0.8)$$
$$= 5.424$$
$$\leq \theta = 7$$

which means (s)he will not watch the movie as the weighted sum lies below the threshold.
$\therefore$ **Option B** is the correct answer.

5. Consider a small training set with the following points in $\mathbb{R}^3$:

Index	Points $[x_0, x, y, z]$	Class
n_1	[1,0,0,0]	Class 0
p_1	[1,0,0,1]	Class 1
p_2	[1,0,1,0]	Class 1
p_3	[1,0,1,1]	Class 1
p_4	[1,1,0,0]	Class 1
p_5	[1,1,0,1]	Class 1
p_6	[1,1,1,0]	Class 1
p_7	[1,1,1,1]	Class 1

Note that there are 8 points which are divided into two classes, Class 0 and Class 1. We are interested in finding the plane which divides the input space into two classes. Starting with the weight vector, $\mathbf{w} = [0, 0, -1, 2]$, apply the perceptron algorithm by going over the points in the following order $[n_1, p_1, p_2, p_3, p_4, p_5, p_6, p_7]$. If needed, repeat in the same order till convergence. After the algorithm converges, what is the value of the weight vector?

A. $\mathbf{w} = [1, 1, 2, 3]$

B. $\mathbf{w} = [-1, 1, 1, 2]$

C. $\mathbf{w} = [-3, -2, -1, -1]$

D. $\mathbf{w} = [-2, -1, -1, 1]$

Solution: You can arrive at the solution by implementing the following pseudo code in python

Algorithm 1: Perceptron Learning Algorithm

$P \leftarrow inputs \quad with \quad label \quad 1;$
$N \leftarrow inputs \quad with \quad label \quad 0;$
Initialize $\mathbf{w} = [0, 0, -1, 2]$;
while !$convergence$ **do**
  for $\mathbf{x} \in [n_1, p_1, p_2, p_3, p_4, p_5, p_6, p_7]$ **do**
    if $\mathbf{x} \in P \quad and \quad \mathbf{w}.\mathbf{x} < 0$ **then**
      $\mathbf{w} = \mathbf{w} + \mathbf{x}$;
    end
    if $\mathbf{x} \in N \quad and \quad \mathbf{w}.\mathbf{x} \geq 0$ **then**
      $\mathbf{w} = \mathbf{w} - \mathbf{x}$;
    end
  end
end
//the algorithm converges when all the inputs are classified correctly
//so after every run of the inner while loop you need to check the number of errors. When the number of errors is 0 the algorithm would converge and you can ouput the value of $\mathbf{w}$.

Option B is the correct answer.

6. A 2-dimensional dataset for 2 classes is given to you. Plot the data and comment whether the 2 classes are linearly separable or not. Note that the top 500 rows are of Class A and the rest 500 are of Class B. You can download the dataset by clicking here. Feel free to use any programming language/plotting tool of your choice. Once you plot the data, answer the following question:
Is the data linearly separable ?

   A. True

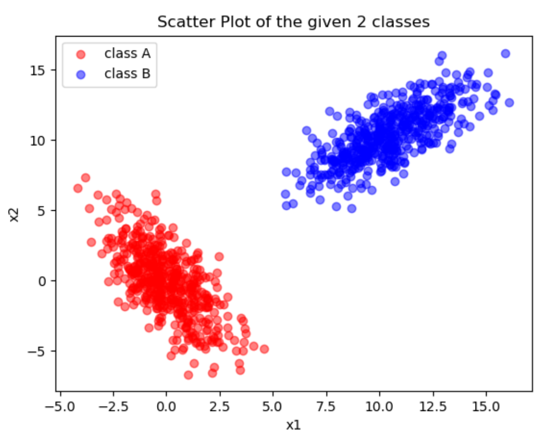
   B. False

Solution: You can plot the data using the following code:

plot.py

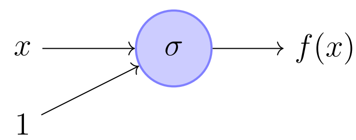
```python
# Import Packages
import numpy as np
import matplotlib.pyplot as plt

# Load data
data = np.loadtxt('Assignment_Q4_data.txt')

# Split the data in two classes
cA = data[:500]
cB = data[500:]

plt.scatter(cA[:,0], cA[:,1], c='red', alpha=0.5, label='class A')
plt.scatter(cB[:,0], cB[:,1], c='blue', alpha=0.5, label='class B')
plt.title("Scatter Plot of the given 2 classes")
plt.xlabel('x1')
plt.ylabel('x2')
plt.legend()
plt.show()
```

As we can see from the scatter plot given below, the two classes are linearly separable.



∴ **Option A** is the correct answer.

7. **Partial derivatives** This question is not based on the material that we have covered so far. However, this is a part of the pre-requisites and will be required for the material that we will cover in the next class. Consider the following function,



$$f(x) = \tfrac{1}{1+e^{-(w \cdot x + b)}}$$

The value L is given by,

$$L = \frac{1}{2}(y - f(x))^2$$

Here, x and y are constants and w and b are parameters that can be modified. In other words, L is a function of w and b.

Derive the partial derivatives, $\frac{\partial L}{\partial w}$ and $\frac{\partial L}{\partial b}$ and choose the correct option.

A. $\frac{\partial L}{\partial w} = (y - f(x))f(x)(1 - f(x))$ and $\frac{\partial L}{\partial b} = (y - f(x))f(x)(1 - f(x))x$

B. $\frac{\partial L}{\partial w} = (y - f(x))(1 - f(x))x$ and $\frac{\partial L}{\partial b} = -(y - f(x))f(x)(1 - f(x))$

C. $\frac{\partial L}{\partial w} = -(y - f(x))f(x)(1 - f(x))x$ and $\frac{\partial L}{\partial b} = -(y - f(x))f(x)(1 - f(x))$

Solution: Let us denote $z = f(x)$. Also, let us derive some elementary derivatives which we then stitch together to create the partial derivatives we seek:

$$L = \frac{1}{2}(y - z)^2$$

$$\Rightarrow \frac{dL}{dz} = -(y - z)$$

Next, if θ is one of the parameters on which $z = f(x)$ depends, using the chain rule of derivatives (and the fact that y is constant), we have:

$$\Rightarrow \frac{\partial L}{\partial \theta} = \frac{dL}{dz}\frac{\partial z}{\partial \theta}$$

$$\Rightarrow \frac{\partial L}{\partial \theta} = -(y - z)\frac{\partial z}{\partial \theta}$$

$$\Rightarrow \frac{\partial L}{\partial \theta} = -(y - f(x))\frac{\partial f(x)}{\partial \theta}$$

For the sigmoid function, let us denote the total input as $wx + b = t$ and find its derivative wrt t:

$$f(x) = \frac{1}{1 + e^{-(wx+b)}} = \frac{1}{1 + e^{-t}}$$

$$\Rightarrow \frac{df(x)}{dt} = \frac{d\frac{1}{1+e^{-t}}}{dt}$$

$$\Rightarrow \frac{df(x)}{dt} = \frac{e^{-t}}{(1 + e^{-t})^2}$$

$$\Rightarrow \frac{df(x)}{dt} = \frac{1}{1 + e^{-t}} - \frac{1}{(1 + e^{-t})^2}$$

$$\Rightarrow \frac{df(x)}{dt} = f(x)(1 - f(x))$$

Again, using the chain rule of probability and for a generic parameter θ:

$$\Rightarrow \frac{\partial f(x)}{\partial \theta} = \frac{df(x)}{dt}\frac{\partial t}{\partial \theta}$$

$$\Rightarrow \frac{\partial f(x)}{\partial \theta} = f(x)(1 - f(x))\frac{\partial t}{\partial \theta}$$

And so stitching it all back together and using the fact that the partial derivative of t wrt w is x and wrt b is 1 we get:

$$\Rightarrow \frac{\partial L}{\partial w} = -(y - f(x))f(x)(1 - f(x))x$$

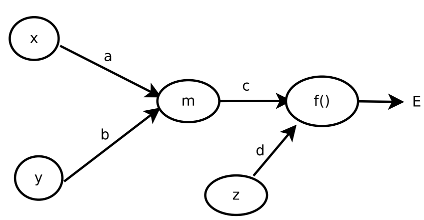
$$\Rightarrow \frac{\partial L}{\partial b} = -(y - f(x))f(x)(1 - f(x))$$

$\therefore$ **Option C** is the correct answer.

8. Consider the function E as given below,

$$E = g(x, y, z) = f(c(ax + by) + dz)$$

Represented as a graph, we have



Here x, y, z are inputs (constants) and a, b, c, d are parameters (variables). m is an intermediate computation and f is some differentiable function. Specifically, let us consider f to be the *tanh* function.

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

Note that here E is a function of a, b, c, d. Compute the following partial derivatives of E with respect to a i.e $\frac{\partial E}{\partial a}$, and choose the correct option.

A. $\frac{\partial E}{\partial a} = (1 - f(c(ax + by) + dz)^2)cx$

B. $\frac{\partial E}{\partial a} = c(1 - f(c(ax + by) + dz)^2)$

C. $\frac{\partial E}{\partial a} = (1 - f(c(ax + by) - dz)^2)cx$

Solution: Let us denote $t = c(ax + by) + dz$. Then by chain rule of probability, for a generic parameter θ we have:

$$\frac{\partial E}{\partial \theta} = \frac{df(t)}{dt}\frac{\partial t}{\partial \theta}$$

We have:

$$f(t) = \frac{e^t - e^{-t}}{e^t + e^{-t}}$$

Using Quotient rule of derivatives we have:

$$\Rightarrow \frac{df(t)}{dt} = \frac{(e^t + e^{-t})^2 - (e^t - e^{-t})^2}{(e^t + e^{-t})^2}$$

$$\Rightarrow \frac{df(t)}{dt} = 1 - f(t)^2$$

Hence, we have:

$$\Rightarrow \frac{\partial E}{\partial a} = (1 - f(c(ax + by) + dz)^2)cx$$

$\therefore$ **Option A** is the correct answer.

9. Keeping the graph discussed in question 8 in mind, find $\frac{\partial E}{\partial b}$ and choose the correct option.

A. $\frac{\partial E}{\partial b} = (1 - f(c(ax + by) + dz))cy$

B. $\frac{\partial E}{\partial b} = (1 - f(c(ax + by) + dz)^2)$

C. $\frac{\partial E}{\partial b} = (1 - f(c(ax + by) + dz)^2)cy$

Solution: Continuing the derivation in part (a) above we have,

$$\Rightarrow \frac{\partial E}{\partial b} = (1 - f(c(ax + by) + dz)^2)cy$$

∴ **Option C** is the correct answer.

10. Keeping the graph discussed in question 8 in mind, find $\frac{\partial E}{\partial c}$ and choose the correct option.

    A. $\frac{\partial E}{\partial c} = (1 - f(c(ax + by) + dz)^2)(ax + by)$

    B. $\frac{\partial E}{\partial c} = (1 - f(c(ax + by) + dz))(ax + by)$

    C. $\frac{\partial E}{\partial c} = (1 - f(c(ax + by) + dz)^2)$

Solution: Continuing the derivation in part (a) above we have,

$$\Rightarrow \frac{\partial E}{\partial c} = (1 - f(c(ax + by) + dz)^2)(ax + by)$$

∴ **Option A** is the correct answer.

11. Keeping the graph discussed in question 8 in mind, find $\frac{\partial E}{\partial d}$ and choose the correct option.

    A. $\frac{\partial E}{\partial d} = 2(1 - f(c(ax + by) + dz)^2)z$

    B. $\frac{\partial E}{\partial d} = (1 - f(c(ax + by) + dz)^2)z$

    C. $\frac{\partial E}{\partial d} = (1 - f(c(ax + by) + dz)^2)$

Solution: Continuing the derivation in part (a) above we have,

$$\Rightarrow \frac{\partial E}{\partial d} = (1 - f(c(ax + by) + dz)^2)z$$

$\therefore$ **Option B** is the correct answer.