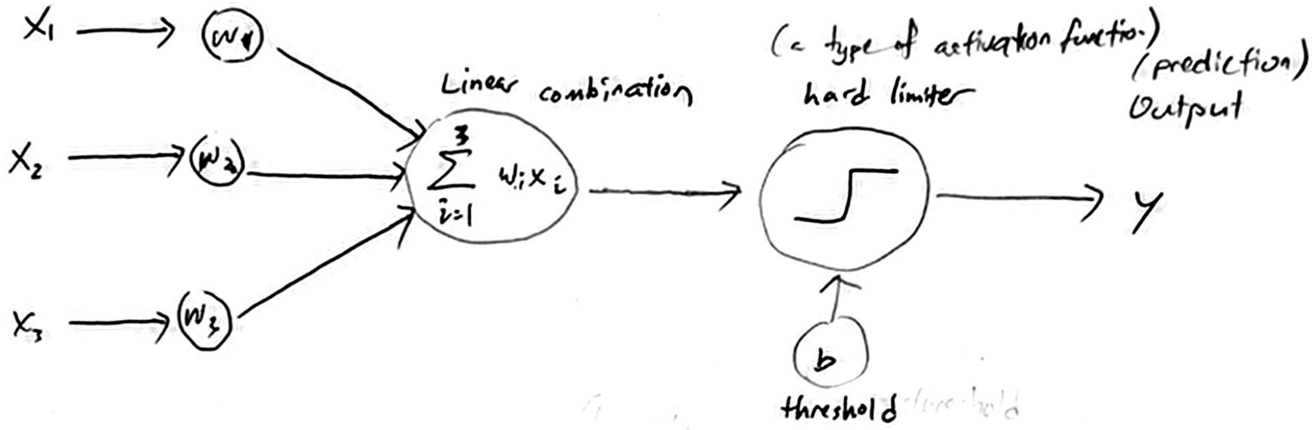
Q1



Single layer perceptron with 3 input nodes.

- In the case of an elementary perceptron, the n-dimensional space is divided by a hyperplane into two decision regions. The hyperplane is defined by linearly separable function.

- The hyperplane equation is as defined below

$$w_1x_1 + w_2x_2 + w_3x_3 = b$$

- If this is generalized, the following equation can be the linear separable function

$$\sum_{i=0}^{3} w_ix_i - b = 0$$

In this case, $w_1x_1 + w_2x_2 + w_3x_3 - b = 0$

If we express the hyperplane equation as matrix

$$(w_1, w_2, w_3)\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = b$$

- The normal vector perpendicular to the hyperplane is (w_1, w_2, w_3)
- and the distance from origin to the hyperplane is $\frac{|b|}{\|w\|_2}$

- If $(w_1, w_2, w_3)\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = b$, then the data is in hyperplane,

- If $(w_1 w_2, w_3)\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} > b$, then the data is located in the direction of upper part of the hyperplane.

- If $(w_1 w_2, w_3)\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} < b$, then the data is located in the direction of below part of the hyperplane.

- For a given dataset, if we can find the (w_1, w_2, w_3) that can satisfy the formula below, then we can say the given data is linear separable, if not, we can say the given data is non-linear seperable.

$$Y \begin{cases} 1, & (w_1, w_2, w_3)\begin{Bmatrix} x_1 \\ x_2 \\ x_3 \end{Bmatrix} \geq b \\ -1, & (w_1, w_2, w_3)\begin{Bmatrix} x_1 \\ x_2 \\ x_3 \end{Bmatrix} \leq b \end{cases}$$

1. For $X_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$

$0 > b$

2. For $X_2 = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}$

$w_2 + w_3 > b$

3. For $X_3 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$

$w_1 + w_2 + w_3 > b$

4. For $X_4 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$

$w_1 + w_3 \leq b$

5. For $X_5 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$

$w_3 \leq b$

6. For $X_6 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$

$w_2 \leq b$

$\therefore$ From the answer of 2 & 3

$w_2 + w_3 > b$ — 2

$w_1 + w_2 + w_3 > b$ — 3

sub 2 into 3

$w_1 > 0$

From the answer of 4 & 5

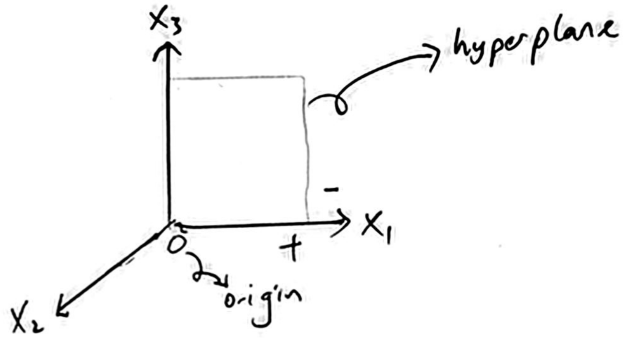
$w_1 + w_3 \leq b$

$w_3 \leq b$

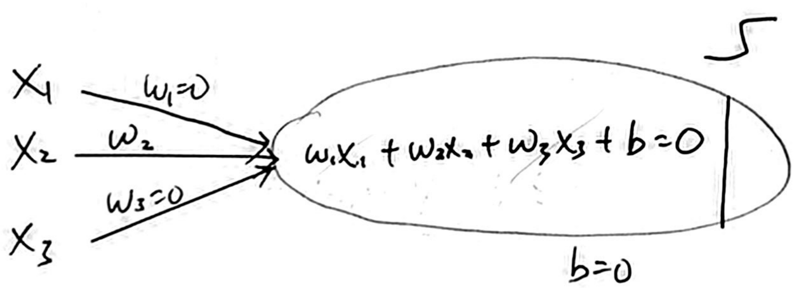
$w_1 < 0$

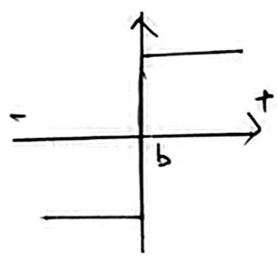
Since the equation of $w_1 > 0$ & $w_1 < 0$ is contradictory to each other, the given dataset cannot form linearly separable hyperplane. Therefore, we can say that the given data is non-linear-seperable.

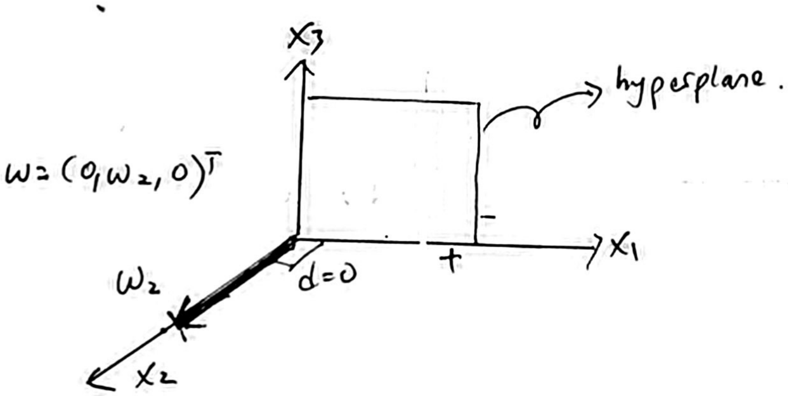
Q2.



- Since the hyperplane (in the figure) is parallel to X_1 & X_2, w_1, w_2 will be equal to 0.
- Also, since the distance of hyperplane from origin is 0, we can know the value of b will be 0.







The linear seperation between perception and hyperplane

Q4 : Objective function $J(w) = \sum_{i=1}^{n} (y_i - \tau(w^T x_i))^2$
(loss function)

$$\frac{\Delta J(w)}{\Delta w_i} = \frac{\Delta\{y_k - \tau(w_1 X_{k1} + w_2 X_{k2} + \dots w_i X_{ki} + w_n X_{kn})\}^2}{\Delta w_i}$$

<u>chain Rule</u>.

$$\frac{\Delta J(w)}{\Delta w_i} = \frac{\Delta\{y_k - \tau(w_1 x_{k1} + \dots w_{x_n} x_{kn})\}^2}{\Delta y_k - \tau(w_1 x_{k1} \dots)} \times \frac{\Delta y_k - \tau(w_1 x_{k1} \dots)}{\Delta (w_1 x_{k1} \dots)} \times \frac{\Delta (w_1 x_{k1} \dots)}{\Delta w_1 x_{k1} \dots}$$

$$\times \frac{\Delta w_1 x_{k1} + \dots + w_i x_{ki} + \dots}{\Delta w_i}$$

$$= 2\{y_k - \tau(w_1 x_{k1} + \dots)\} \times (-1) \times \tau'(w_1 x_{k1} + \dots) \times x_{ki}$$

$$\therefore \frac{\Delta J(w)}{\Delta w_i} = \sum_{x_k \in Y} -2x_{ki} (y_k - \tau(w_1 x_{k1} + \dots)) \tau'(w_1 x_{k1} + \dots)$$

Delta rule
(델타 규칙)

$$w_i = w_i - \rho \frac{\partial J(w)}{\partial w_i} \quad ; \rho : \text{learning rate}$$

$$w = w - \rho \sum_{x_k \in Y} -2x_{ki}(y_k - \tau(w_1 x_{k1} + \dots)) \tau'(w_1 x_{k1} + \dots)$$

$$w = w + \rho \sum_{x_k \in Y} y_k x_k$$

※ proven.

(Q3)

$$y = (w_0 + w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_8, w_9)\begin{pmatrix} z_0 = 1 \\ z_1 \\ z_2 \\ z_3 \\ z_1 z_2 \\ z_2 z_3 \\ z_1 z_3 \\ z_1^2 \\ z_2^2 \\ z_3^2 \end{pmatrix}$$

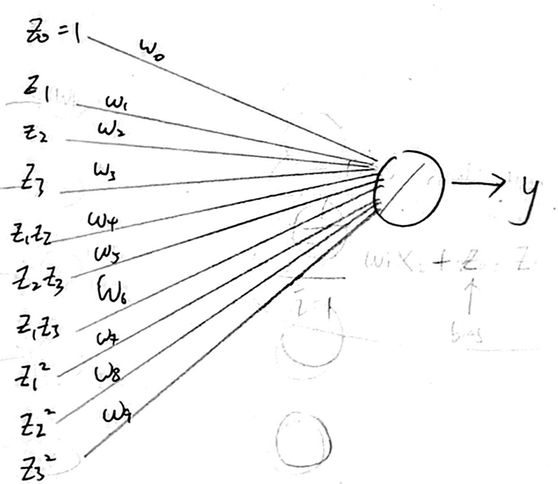
$$y = wz$$

$$y = w_0 + w_1 z_1 + w_2 z_2 + w_3 z_3 + w_4 z_1 z_2 + w_5 z_2 z_3 + w_6 z_1 z_3 + w_7 z_1^2 + w_8 z_2^2 + w_9 z_3^2$$

The dataset can be transformed as shown in table below.

x_k	z_0	z_1	z_2	z_3	$z_1 z_2$	$z_2 z_3$	$z_1 z_3$	z_1^2	z_2^2	z_3^3	Y
X_1	1	0	0	0	0	0	0	0	0	0	1
X_2	1	0	1	1	0	1	0	0	1	1	-1
X_3	1	1	1	1	1	1	1	1	1	1	1
X_4	1	1	0	1	0	0	1	1	0	1	-1
X_5	1	0	0	1	0	0	0	0	0	1	1
X_6	1	0	1	0	0	0	0	0	1	0	-1

If transform the data from original space to feature space, then this data can be classified (



→ According to the gradient descent, we initialize w.

→ Then, we update the w as:

$$w_{t+1} = w_t + Lr \cdot (Y - \hat{Y}) X_i$$

→ After the initialization of weight, update of weight, and result is optimized (shown in ipynb file)