Assignment 7

ECE 4200

{ Name: Zihao Lin
Netid: zl755 }

Problem #1:

- $\vec{x} \in \mathbb{R}^d$; all entries non-zero
- $w \in \mathbb{R}^{m \times d}$
- $\vec{y} = \max\{0, w\vec{x}\}$
- w follows $N(0,1)$ distribution.

1. $\vec{y} = \max\{0, w\vec{x}\}$

$$\vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} \Big\} d \qquad w\vec{x} = \begin{bmatrix} w_{11} \cdots w_{1d} \\ \vdots \\ w_{m1} \cdots w_{md} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$$

m

$$w = \begin{bmatrix} w_{11} \cdots w_{1d} \\ \vdots \qquad \vdots \\ w_{m1} \cdots w_{md} \end{bmatrix} \Big\} m \qquad = \begin{bmatrix} x_1 w_{11} + \cdots + x_d w_{1d} \\ \vdots \\ x_1 w_{m1} + \cdots = w_{md} x_d \end{bmatrix}$$

$\underbrace{\qquad\qquad}_{d}$

$$\vec{y} = \max\{0, W\vec{x}\}$$

$$= \max\left\{ \vec{0}, \begin{bmatrix} x_1 w_{11} + \cdots + x_d w_{1d} \\ \vdots \\ x_1 w_{m1} + \cdots + w_{md} x_d \end{bmatrix} \right\}$$

Component-wise max function

Each entry of W is independent, distributed according to $N(0,1)$:

$$\text{let } W\vec{x} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} = \begin{bmatrix} x_1 w_{11} + \cdots + x_d w_{1d} \\ \vdots \\ x_1 w_{m1} + \cdots + w_{md} x_d \end{bmatrix}$$

since elements of w are normally distributed and independent:

$$E(a_1) = E(x_1 w_{11} + \cdots + x_d w_{1d}) \quad \text{(linearity)}$$

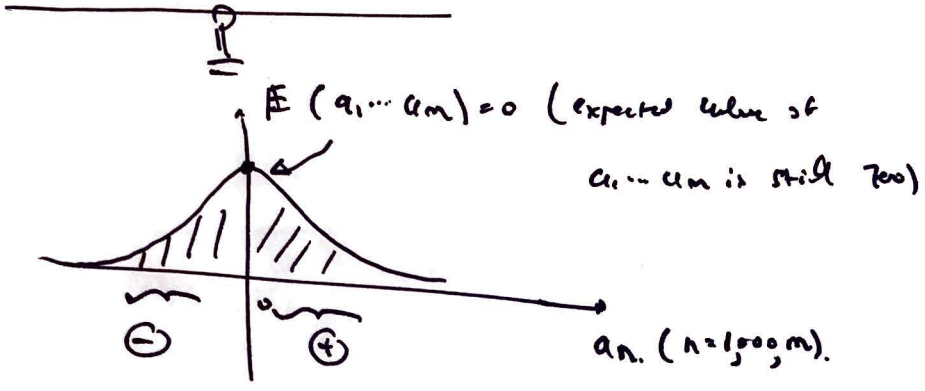
$$= E(x_1 w_{11}) + \cdots + E(x_d w_{1d})$$

$$= x_1 \underbrace{E(w_{11})}_{\text{mean is 0}} + \cdots + x_d \underbrace{E(w_{1d})}_{\text{mean is 0}}$$

$$= 0$$

Similarly, $E(a_2) = \cdots = E(a_m) = 0$

Thus, we see that $a_1 \cdots a_m$ are again distributed with mean of 0. In other words, we expected that half of $a_1 \cdots a_m$ are greater than 0 and half of $a_1 \cdots a_m$ are less than 0.



Since $\vec{y} = \max\{0, w\vec{x}\}$ and half of $a_1 \cdots a_m$ is expected to be greater than 0 and half is expected to be less than 0, around half of entries in $\vec{y}$ are 0.

$\therefore$ Thus, $m/2$ entries in $\vec{y}$ are expected to be non-zero.

2. $\|\vec{x}\|_2 = r^2 = x_1^2 + \cdots + x_d^2$

$$w\vec{x} = \begin{bmatrix} a_1 \\ \vdots \\ a_m \end{bmatrix} = \begin{bmatrix} x_1 w_{11} + \cdots + x_d w_{1d} \\ \vdots \\ x_1 w_{m1} + \cdots + x_d w_{md} \end{bmatrix}$$

$\Rightarrow$ the mean of $a_1 \cdots a_m$ will be zero as shown before.

Note that:

* $X \sim N(\mu_x, \sigma_x^2)$, $Y \sim N(\mu_y, \sigma_y^2)$, $Z = X + Y \rightarrow Z \sim N(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$

* $E(c \cdot x) = c \cdot E(X)$
  $Var(c \cdot X) = c^2 \cdot Var(X)$
  $\sigma^2(c \cdot x) = c^2 \sigma^2(X)$

$q_n = x_1 w_{n1} + \dots + x_d w_{nd}$

- $x_1 w_{n1} \sim N(0, x_1^2 \cdot 1) = N(0, x_1^2)$
- $x_2 w_{n2} \sim N(0, x_2^2 \cdot 1) = N(0, x_2^2)$
- $\vdots$
- $x_d w_{nd} \sim N(0, x_d^2)$ $\quad |x_i|^2 = \sigma^2$

Thus, $Z \sim N(0 + \dots + 0, x_1^2 + \dots + x_d^2)$

$\neq N(0, \sigma^2)$

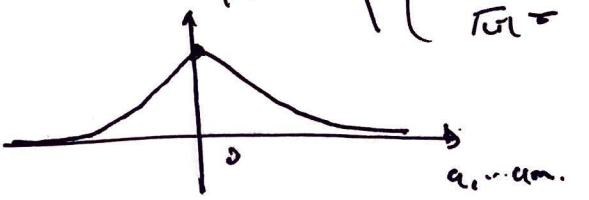
$\therefore$ the entries of w_i are distributed differently to: $N(0, \sigma^2)$.

3. Each entry in $\vec{y}$:

$$\vec{y} = \begin{bmatrix} \max(0, a_1) \\ \max(0, a_2) \\ \vdots \\ \max(0, a_m) \end{bmatrix} = \begin{bmatrix} a_1' \\ a_2' \\ \vdots \\ a_m' \end{bmatrix}$$

$\Rightarrow$ $a_1 \sim a_m \sim N(0, \sigma^2)$

probability $\left(\frac{e^{-a_n^2/2\sigma^2}}{\sqrt{2\pi}\sigma} \right)$



$$E(a_n') = \int_{-\infty}^{\infty} \max(0, a_n) p(a_n) \, da_n$$

(< 0 part doesn't contribute anything)

$$= \int_0^{\infty} \max(0, a_n) \, p(a_n) \, da_n$$

$$= \int_0^{\infty} a_n \cdot \frac{e^{-a_n^2/2\sigma^2}}{\sqrt{2\pi}\sigma} \, da_n$$

$a_n \sim N(0, \sigma^2) \Rightarrow \frac{e^{-a_n^2/2\sigma^2}}{\sqrt{2\pi}\sigma}$

$$= \frac{\sigma}{\sqrt{2\pi}}$$

∴ The expected value / mean of every entry in $\vec{y}$ is $\frac{\sigma}{\sqrt{2\pi}}$.

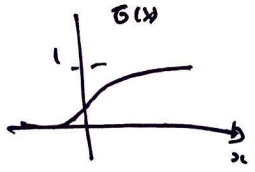
Problem 2:

$m=2,\ d=2.$

$$W=\begin{bmatrix}1 & 2\\ -2 & 3\end{bmatrix},\quad \vec{x}=\begin{bmatrix}2\\ -3\end{bmatrix}.$$

$$\text{let } L=\max\left\{\sigma(w_{(1)}\vec{x}),\sigma(w_{(2)}\vec{x})\right\}$$

$$\sigma(x)=\frac{1}{1+e^{-x}}.$$

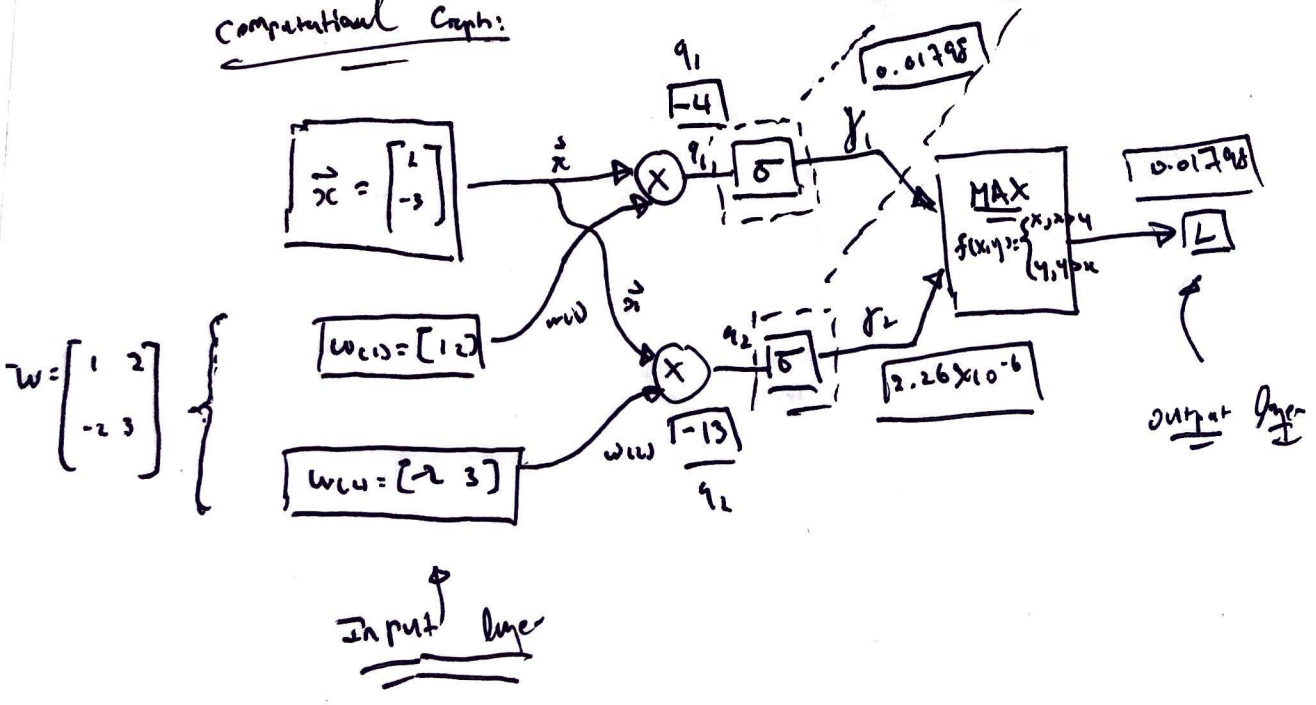


Ans:

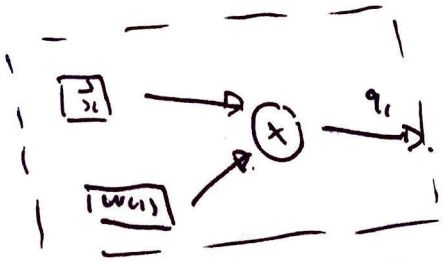
Row vector:

$$\begin{cases} w_{(1)}=\begin{bmatrix}1 & 2\end{bmatrix}=\begin{bmatrix}1\\2\end{bmatrix}^T \\ w_{(2)}=\begin{bmatrix}-2 & 3\end{bmatrix}=\begin{bmatrix}-2\\3\end{bmatrix}^T \end{cases}$$

Computational Graph:



Computation of gradients:



$\vec{x_i} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

$w_{(1)} = [w_1^{(1)} \; w_2^{(1)}]$

$q_1 = w_{(1)} \cdot \vec{x_i}$

$$q_1 = [w_1^{(1)} \; w_2^{(1)}] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = q_1(\vec{x}, w_{(1)})$$

$$= [x_1 w_1^{(1)} + x_2 w_2^{(1)}] = q_1$$

Jacobian:

1×1 matrix → treat as scalar.

$$\Rightarrow \boxed{\frac{\partial q_{1,1}(\vec{x}, w_{(1)})}{\partial x_j} = w_j^{(1)}}$$

$$w_1^{(1)} = \frac{\partial q_{1,1}}{\partial x_1}$$

$$w_2^{(1)} = \frac{\partial q_{1,1}}{\partial x_2}$$

Do row

$$\therefore \; w_1^{(1)} = 1 = \frac{\partial q_{1,1}}{\partial x_1}$$

$$w_2^{(1)} = 2 = \frac{\partial q_{1,1}}{\partial x_2}$$

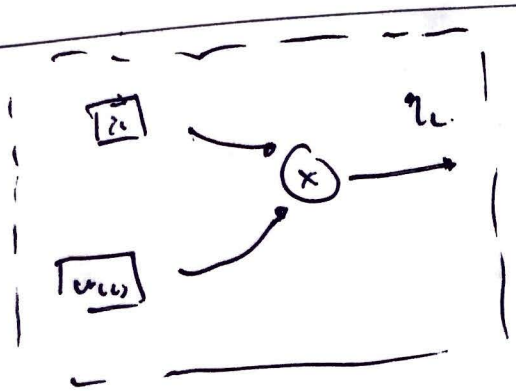
$$\Rightarrow \boxed{\frac{\partial q_{1,1}(\vec{x}, w_{(1)})}{\partial w_j^{(1)}} = x_j}$$

$$x_1 = \frac{\partial q_{1,1}}{\partial w_1^{(1)}}$$

$$x_2 = \frac{\partial q_{1,1}}{\partial w_2^{(1)}}$$

$$\therefore \; w_1 = 2 = \frac{\partial q_{1,1}}{\partial w_1^{(1)}}$$

$$x_2 = -3 = \frac{\partial q_{1,1}}{\partial w_2^{(1)}}$$



$\vec{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

$w_{(2)} = [w_1^{(2)} \; w_2^{(2)}]$

$$q_2 = [w_1^{(2)} \; w_2^{(2)}] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$= [x_1 w_1^{(2)} + x_2 w_2^{(2)}]$$

Jacobian:

$$\frac{\partial q_{2,1}(\bar{x}, w_{(2)})}{\partial x_j} = w_j^{(2)}$$

$w_1^{(2)} = \frac{\partial q_{2,1}}{\partial x_1}$ $\quad w_2^{(2)} = \frac{\partial q_{2,1}}{\partial x_2}$

$w_1^{(2)} = -2 = \frac{\partial q_{2,1}}{\partial x_1}$

$\therefore w_2^{(2)} = 3 = \frac{\partial q_{2,1}}{\partial x_2}$

$$\frac{\partial q_{2,1}(\bar{x}, w_{(2)})}{\partial w_j^{(2)}} = x_j$$

$x_1 = \frac{\partial q_{2,1}}{\partial w_1^{(2)}}$ $\quad x_2 = \frac{\partial q_{2,1}}{\partial w_2^{(2)}}$

$x_1 = 2 = \frac{\partial q_{2,1}}{\partial w_1^{(2)}}$

$\therefore x_2 = -3 = \frac{\partial q_{2,1}}{\partial w_2^{(2)}}$

$q_1 \rightarrow [\sigma] \rightarrow \delta_1$

$q_1 = -4$

$\delta_1 = \sigma(q_1) = \frac{1}{1+e^{-q_1}}$

$\frac{\partial \delta_1}{\partial q_1} = \frac{\partial}{\partial q_1}[\sigma(q_1)] = \sigma(q_1)(1-\sigma(q_1))$

$\therefore \frac{\partial \delta_1}{\partial q_1} = \sigma(-4)(1-\sigma(-4))$

$= 0.01766$

$q_2 \rightarrow [\sigma] \rightarrow \delta_2$ $\quad q_1 = -13$

$\frac{\partial \delta_2}{\partial q_2} = \sigma(q_2)(1-\sigma(q_2))$

$\therefore \frac{\partial \delta_2}{\partial q_2} = \sigma(-13)(1-\sigma(-13))$

$= 2.26 \times 10^{-6}$

$\delta_1 (-4)$

$(-13)\ \delta_2$ → [Mux] → L.

Since $\delta_1 > \delta_2$, gradient of 0.01766 passes through mux gate.

- δ_1 gradient passes if $\delta_1 > \delta_2$.
- δ_2 gradient passes if $\delta_2 > \delta_1$

Problem #3:

$$\hat{y}(z_1, z_2) = \frac{e^{z_1}}{e^{z_1} + e^{z_2}}$$

Cross-entropy loss:

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

Goal: Show that: ① $\frac{\partial L(y,\hat{y})}{\partial z_1} = \hat{y} - y$; ② $\frac{\partial L(y,\hat{y})}{\partial z_2} = y - \hat{y}$

---

① $$\frac{\partial L}{\partial z_1} = \frac{\partial}{\partial z_1}\left[-y \log(\hat{y}) - (1-y)\log(1-\hat{y})\right]$$

(y is not function of z_1 but $\hat{y}$ is

$$= -y \frac{\partial}{\partial z_1} \log(\hat{y}) - (1-y) \frac{\partial}{\partial z_1} \log(1-\hat{y})$$

$$= -y \frac{\partial}{\partial z_1}\left[\log\left[\frac{e^{z_1}}{e^{z_1} + e^{z_2}}\right]\right] - (1-y) \frac{\partial}{\partial z_1}\left[\log\left(\frac{e^{z_2}}{e^{z_1} + e^{z_2}}\right)\right]$$

$$= -y \frac{\partial}{\partial z_1}\left[\log e^{z_1} - \log(e^{z_1} + e^{z_2})\right] - (1-y) \frac{\partial}{\partial z_1}\left[\log e^{z_2} - \log(e^{z_1} + e^{z_2})\right]$$

$$= -y \frac{\partial}{\partial z_1}\left[z_1 - \log(e^{z_1} + e^{z_2})\right] - (1-y) \frac{\partial}{\partial z_1}\left[z_2 - \log(e^{z_1} + e^{z_2})\right]$$

$$= -y\left[1 - \frac{e^{z_1}}{e^{z_1}+e^{z_2}}\right] - (1-y)\left[0 - \frac{e^{z_1}}{e^{z_1}+e^{z_2}}\right] \quad (\hat{y})$$

$$= -y + y\cdot\hat{y} + (1-y)\hat{y}$$

$$= -y + y\cdot\hat{y} + \hat{y} - y\hat{y} = \boxed{\hat{y} - y} \quad \therefore \text{Q=P}$$

(2): $$\frac{\partial L}{\partial z_1} = \frac{\partial}{\partial z_1}\left[-y\log(\hat{y}) - (1-y)\log(1-\hat{y})\right]$$

$$= -y\frac{\partial}{\partial z_1}\log(\hat{y}) - (1-y)\frac{\partial}{\partial z_1}\log(1-\hat{y})$$

$$= -y\left[\frac{\partial}{\partial z_1}\log\left(\frac{e^{z_1}}{e^{z_1}+e^{z_2}}\right)\right] - (1-y)\frac{\partial}{\partial z_1}\left[\frac{e^{z_2}}{e^{z_1}+e^{z_2}}\right]$$

$$= -y\frac{\partial}{\partial z_1}\left[\log e^{z_1} - \log\left(e^{z_1}+e^{z_2}\right)\right] - (1-y)\frac{\partial}{\partial z_1}\left[\log e^{z_2} - \log\left(e^{z_1}+e^{z_2}\right)\right]$$

$$= -y\left[0 - \frac{e^{z_1}}{e^{z_1}+e^{z_2}}\right] - (1-y)\left[0 - \frac{e^{z_1}}{e^{z_1}+e^{z_2}}\right]$$

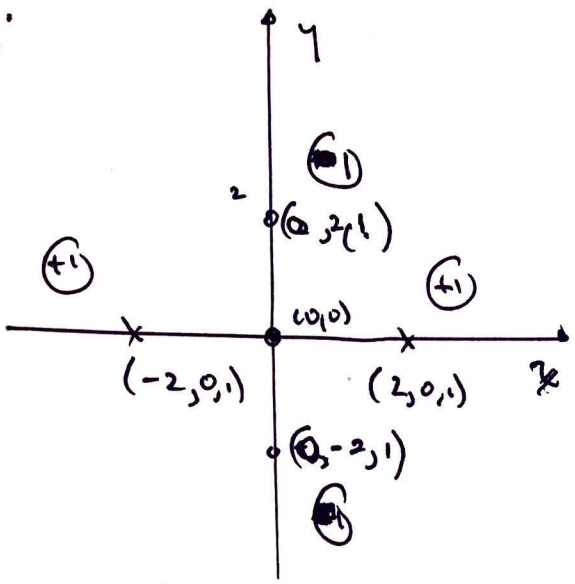
$$= y\frac{e^{z_1}}{e^{z_1}+e^{z_2}} - (1-y)\left(\frac{e^{z_1}}{e^{z_1}+e^{z_2}}\right) \quad (\hat{y})$$

$$= y(1-\hat{y}) - (1-y)\hat{y}$$

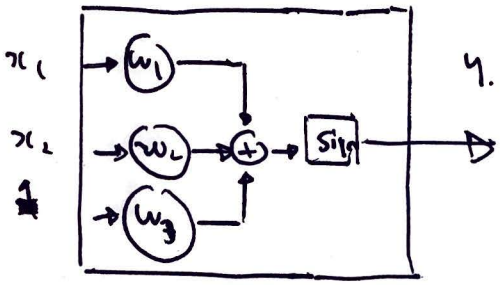
$$= y - y\hat{y} - \hat{y} + y\hat{y} = \boxed{y - \hat{y}} \quad \therefore \text{d=P}$$

Problem # 4.

1.



Unit:

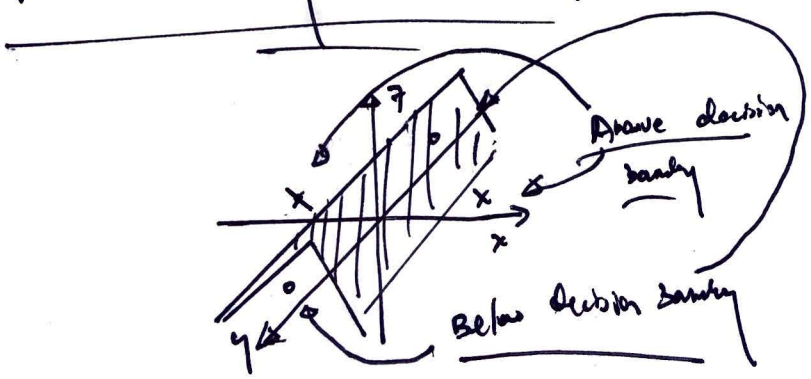


$$y = \text{sign}(x_1 w_1 + x_2 w_2 + w_3)$$

The goal is to correctly classify all four points.

we can do this by creating decision boundaries in $\mathbb{R}^3$ space

we have the following decision boundary:

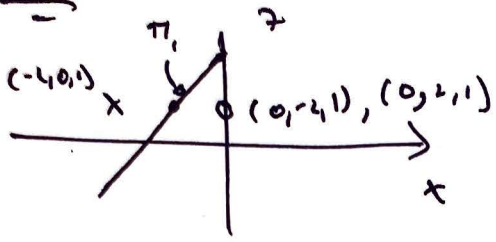


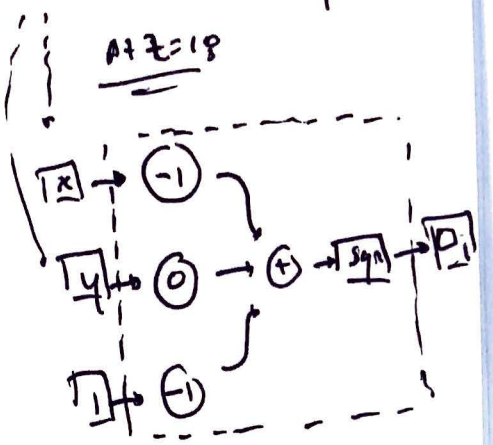
This decision boundary is composed of two planes:

(x,y) coord of Points to classify

plane π_1: $-x+y-2=0$

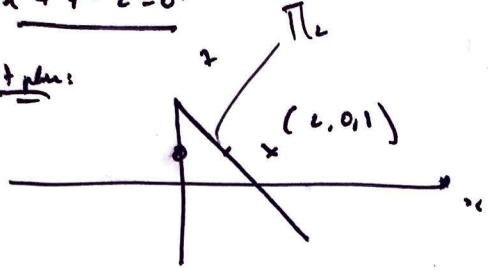
x-z plane:

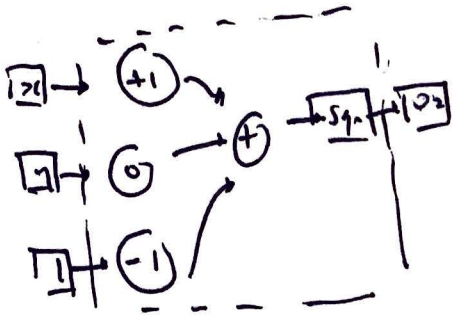




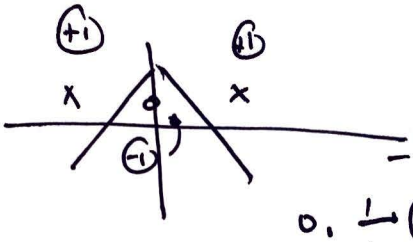
plane π_2: $x+y-2=0$

x-z plane:

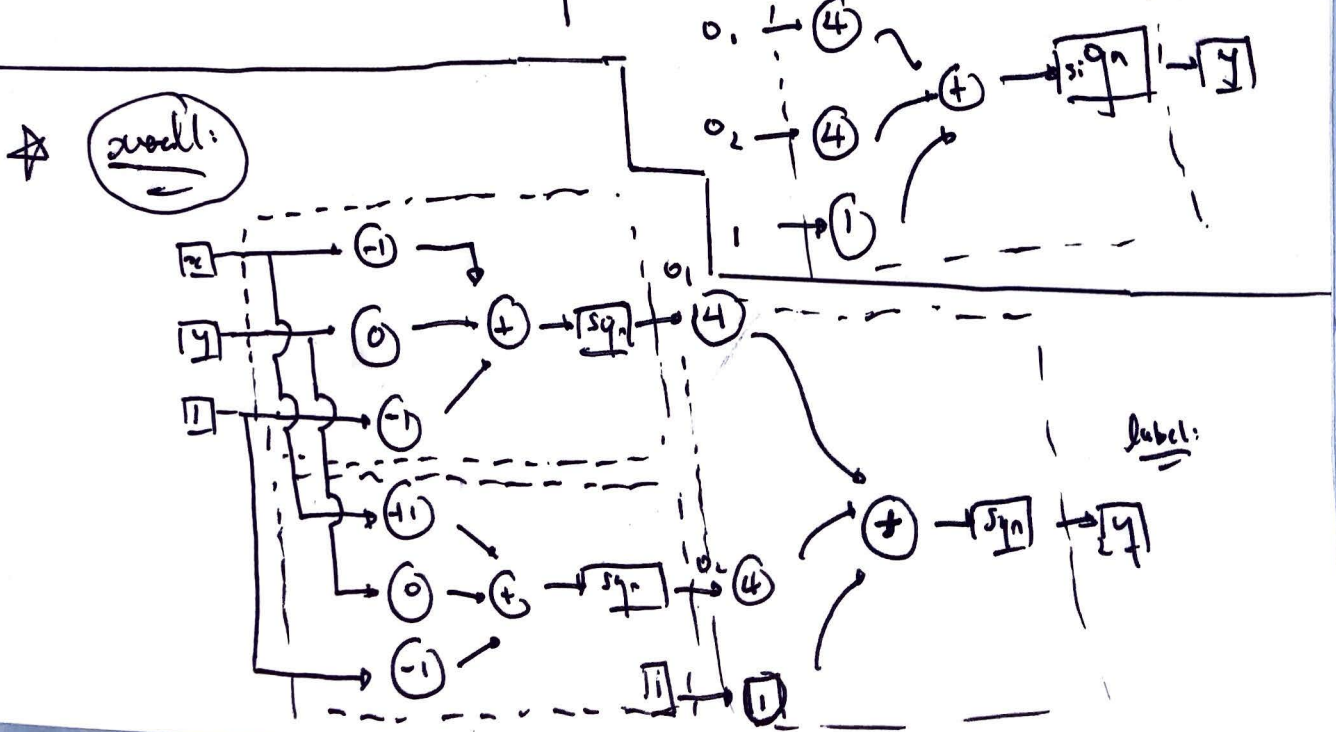




combined:



overall:



Verification:

(-2, 0, 1):

$o_1 = sgn(-1 \times -2 + 0 \times 0 + 1 \times 1) = 1$

$o_2 = sgn(+1 \times -2 + 0 \times 0 + 1 \times (-1)) = -1$

$y = sgn(4 \times 1 + 4 \times -1 + 1 \times 1) = 1$ ✓

(2, 0, 1):

$o_1 = sgn(-1 \times 2 + 0 \times 0 - 1) = -1$

$o_2 = sgn(1 \times 2 + 0 \times 0 - 1) = 1$

$y = sgn(4 \times -1 + 4 \times 1 + 1) = 1$ ✓

(0, 2, 1):

$o_1 = sgn(-1 \times 0 + 2 \times 0 - 1) = -1$

$o_2 = sgn(1 \times 0 + 2 \times 0 - 1) = -1$

$y = sgn(4 \times -1 + 4 \times -1 + 1) = -1$ ✓

$(0, -2, 1)$:

$o_1 = \text{sgn}(-1 \times 0 + (-1) \times 0 - 1) = -1$

$o_2 = \text{sgn}(+1 \times 0 + (-1) \times 0 - 1) = -1$

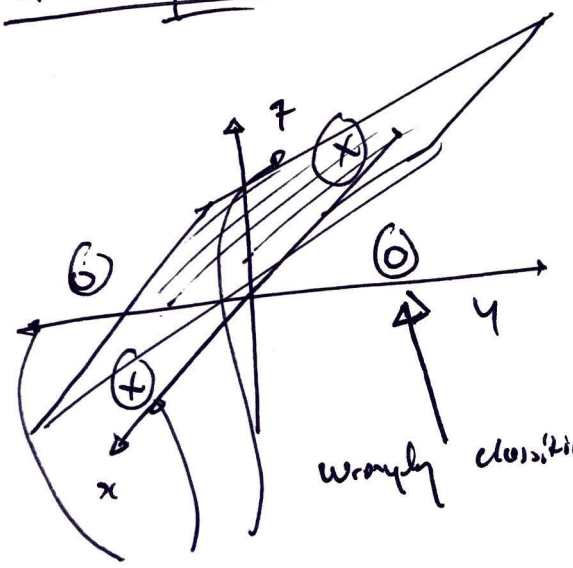
$y = \text{sgn}(-4 - 4 + 1) = -1$ ✓

∴ All correctly classified.

---

2. If we use one unit, the points cannot be all classified correctly because the cluster points are not linearly separable in $\mathbb{R}^3$ space with only one decision boundary. One unit corresponds to one decision boundary.

Not linearly separable:



⇒ It is impossible to create a single plane that correctly separates dataset into two distinct categories

⇒ Not linearly separable with one plane/one unit in $\mathbb{R}^3$ space