Sol^ ① Let L be the multi-class cross entropy loss. which is given by

$$L = -y^T \log \hat{y}$$

$$= -\begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} \log \hat{y}_1 \\ \log \hat{y}_2 \end{bmatrix} \quad \left[\because y = \begin{bmatrix} 1 & 0 \end{bmatrix}^T \text{(given)}\right]$$

$$= -(\log \hat{y}_1 + 0)$$

$$= -\log \hat{y}_1 = -\log\left(\frac{e^{b_1}}{e^{b_1} + e^{b_2}}\right)$$

.5

And the update equation for U is given by:

$$U = U - \eta \frac{\partial L}{\partial U} \quad \text{—①}$$

where η is the learning rate and $\frac{\partial L}{\partial U}$ is the derivati of the loss funct^ wrt U.

Now, computing $\frac{\partial L}{\partial U}$

$$\frac{\partial L}{\partial U} = \left(\frac{\partial L}{\partial a}\right) \frac{\partial a}{\partial U}$$

where $\frac{\partial L}{\partial a}$ is given by

$$\frac{\partial L}{\partial a} = \left(\frac{\partial L}{\partial b1}\right) \cdot \frac{\partial b1}{\partial a} + \left(\frac{\partial L}{\partial b2}\right) \cdot \frac{\partial b2}{\partial a}$$

1

$$= \left(\frac{-1}{e^{b_1}/(e^{b_1}+e^{b_2})} \cdot \frac{(e^{b_1}+e^{b_2})e^{b_1} - e^{b_1}\cdot e^{b_1}}{(e^{b_1}+e^{b_2})^2} \right) \cdot U_1$$

$$+ \left(\frac{-1}{e^{b_1}/(e^{b_1}+e^{b_2})} \cdot \frac{(e^{b_1}+e^{b_2})\cdot 0 - e^{b_1}\cdot e^{b_2}}{(e^{b_1}+e^{b_2})^2} \right) \cdot U_2$$

$$\left[\because \; b_1 = U_1 z = U_1 a \quad \text{and} \quad b_2 = U_2 z = U_2 a \right.$$
$$\left. \Rightarrow \frac{\partial b_1}{\partial a} = U_1 \qquad \Rightarrow \frac{\partial b_2}{\partial a} = U_2 \right]$$

$$= -\left(\frac{\cancel{(e^{b_1}+e^{b_2})}}{e^{b_1}} \cdot \frac{\cancel{e^{2b_1}} + e^{(b_1+b_2)} - \cancel{e^{2b_1}}}{(e^{b_1}+e^{b_2})^{\cancel{2}}} \cdot U_1 \right.$$
$$\left. + \frac{\cancel{(e^{b_1}+e^{b_2})}}{e^{b_1}} \cdot \frac{(-e^{(b_1+b_2)})}{(e^{b_1}+e^{b_2})^{2}} \cdot U_2 \right)$$

$$= -\frac{e^{b_1+b_2}}{e^{b_1}(e^{b_1}+e^{b_2})} \left[U_1 - U_2 \right]$$

$$= -\frac{e^{b_2}}{e^{b_1}+e^{b_2}} \left[U_1 - U_2 \right]$$

$$= -\hat{y}_2 \left[U_1 - U_2 \right]$$

$= \hat{y}_2 [v_1 - v_2]$

$\Rightarrow \frac{\partial L}{\partial v} = (\hat{y}_2 \cdot (v_2 - v_1)) \cdot x$ $\quad \left[\because a = vx \Rightarrow \frac{\partial a}{\partial v} = x \right]$

Substituting this value in (1) we get

$$v = v - \eta \hat{y}_2 \cdot (v_2 - v_1) \cdot x$$

which is the required update equation.

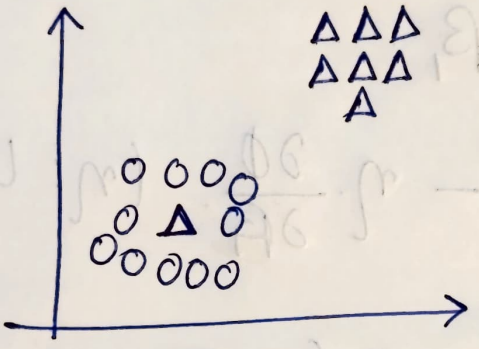
Q2.) (a) The condition is that the training dataset D is linearly separable.
Mathematically, D is linearly separable if there exists a positive constant γ and a weight vector β such that

$$(\beta^T x_i + \beta_0) y_i > \gamma \quad \forall \; 1 < i < n$$

0.5

(b) When the dataset is not linearly separable then gradient descent does not converge for Rosenblatt's perceptron.

eg:



→ Not linearly separable.

0.5

(c) Given, $Z = \text{sign}(b(\beta_1 x + \beta_2 x))$

b is a known constant and not learnable

β_1, β_2 are weights and learnable parameters.

If $Z = 1 \Rightarrow b(\beta_1 x + \beta_2 x) > 0 \Rightarrow y = 1$

If $Z = -1 \Rightarrow b(\beta_1 x + \beta_2 x) < 0 \Rightarrow y = -1$.

If point x is misclassified

i.e $y = 1$ & $b(\beta_1 x + \beta_2 x) < 0$

$y = -1$ & $b(\beta_1 x + \beta_2 x) > 0$.

∴ If x is misclassified $-y(b(\beta_1 x + \beta_2 x)) > 0$

Loss function: $\phi(\beta_1, \beta_2) = -yb(\beta_1 x + \beta_2 x) > 0$
for misclassified points. 0.5

If M: All misclassified points

then
$$\phi(\beta_1, \beta_2) = -\sum_{i \in M} y_i b(\beta_1 x_i + \beta_2 x_i)$$

(d) $$\frac{\partial \phi}{\partial \beta_1} = -\sum_{i \in M} y_i \cdot b \cdot x_i = -b \cdot \sum_{i \in M} y_i \cdot x_i$$

Update rule for β_1:

$$\beta_1^{new} \leftarrow \beta_1^{old} - \eta \cdot \frac{\partial \phi}{\partial \beta}$$ (η: learning rates)

$$\Rightarrow \beta_1^{new} \leftarrow \beta_1^{old} - \eta \cdot \left(-b \sum_{i \in M} \cdot y_i \cdot x_i\right)$$

$$\Rightarrow \beta_1^{new} \leftarrow \beta_1^{old} + \eta \cdot b \cdot \sum_{i \in M} \cdot y_i \cdot x_i$$ 0.5

Q3.

$D = \{(x, y)\} = \{(0,2), (2,1), (1,-1), (-1,-1), (3,-1)\}$

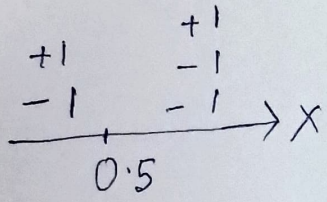
a) $h_1(x) = \text{sign}(y - F_0(x))$

For (x, y) in $D \rightarrow$

$h_1(x) = \{\text{sign}(2-0.5), \text{sign}(1-0.5), \text{sign}(-1-0.5), \text{sign}(-1-0.5), \text{sign}(-1-0.5)\}$

$= \{+1, +1, -1, -1, -1\}$ [updated labels]

b)



Cut at 0.5

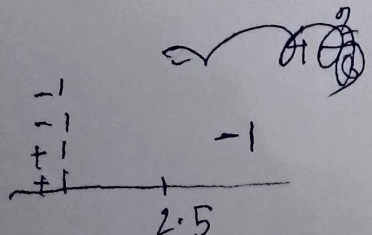
mean at left region 0

right region $= -\frac{1}{3}$

$\text{loss} = |0-1| + |0+1| + |-\frac{1}{3}-1| + |-\frac{1}{3}+1| + |-\frac{1}{3}+1|$

$= |1-0| + |1-0| + |-1+\frac{1}{3}| + |-1+\frac{1}{3}| + |1+\frac{1}{3}|$

$= 1 + 1 + \frac{2}{3} + \frac{2}{3} + \frac{4}{3} = 4\frac{2}{3}$



Mean at left region 0

right region $= -1$

cut at 2.5

$\text{loss} = |1-0| + |-1-0| + |-1-0| + |-1-0| + |-1+1|$

$= 4$

Split at 2.5 is better.

$h_2(x) = +1$ if $x \leq 2.5$

$= -1$ $x > 2.5$

c) $h_1(3.5) = -1$ .5

DATE: __/__/____
PAGE: _____

Q4) Let x_1, x_2, x_3, x_4 be input values, y_1, y_2 be hidden layer output, w_{ij} is the weight between node i and node j and w_{hi} is the weight between hidden layer and output layer. The dropout vectors are $[1,0,0,0]$ and $[0,1]$.

$$\begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \tilde{x}_3 \\ \tilde{x}_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \circ \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} x_1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$y_1 = w_{11} x_1 + b_{11}$$
$$y_2 = w_{12} x_1 + b_{12}$$

.5

$$\begin{bmatrix} \tilde{y}_1 \\ \tilde{y}_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \circ \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 0 \\ w_{12} x_1 + b_{12} \end{bmatrix}$$

$$\text{Output} = w_{h2}\left[w_{12} x + b_{12}\right] + b_{h2}$$

$$= w_{h2} \cdot w_{12} \cdot x + w_{h2} \cdot b_{12} + b_{h2}$$

.5