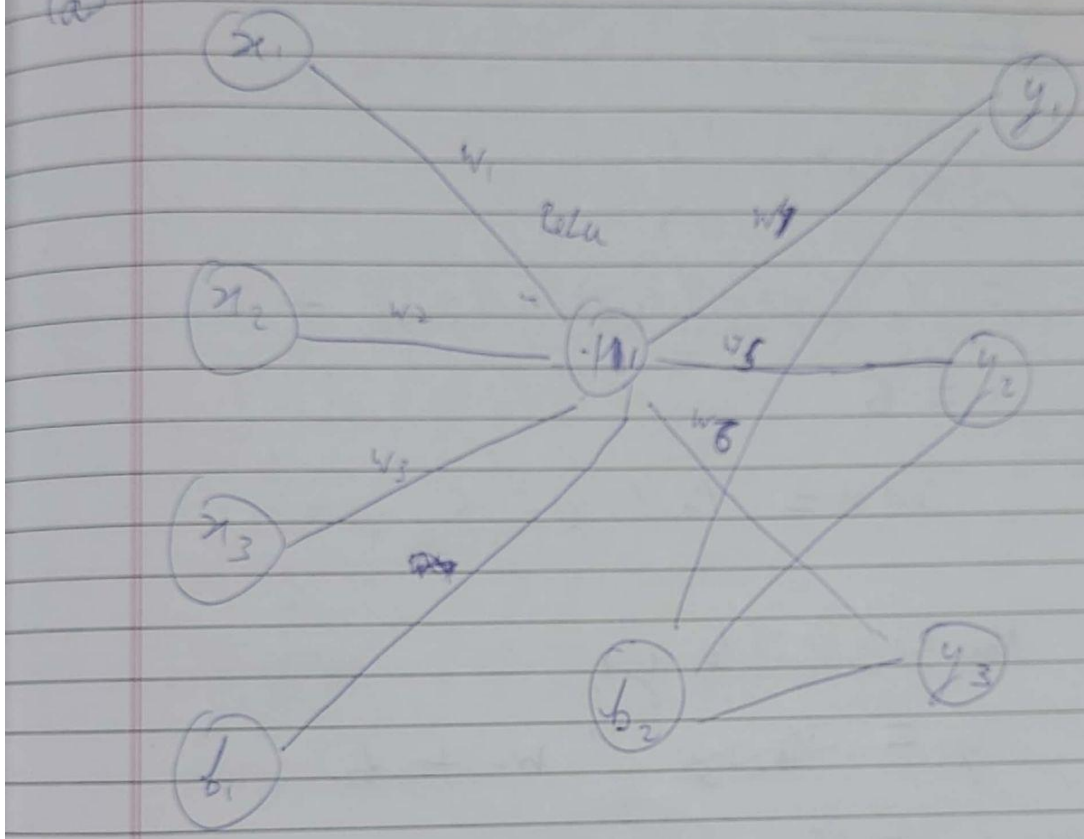
Section - C

1) These are the optimal hyper parameters for this model
   best_params = {'batch_size': 64, 'hidden_layer_sizes': (25, 25), 'learning_rate_init': 0.01, 'max_iter': 20}
2) The best activation function was =ReLu and it gave the accuracy of 65%
3) Possible reason for misclassification is due to overfitting of model on training data. Another reason could be the possible noise in the data as several samples had multiple numbers in the image.

Section - A

(a)



$H = w_1 x_1 + w_2 x_2 + w_3 x_3 + b_1$

Activation Function = ReLu

Out $H_{01} = \max(0, H)$

$\alpha = 0.01$ (Given)

Let $w_1 = w_2 = w_3 = w_4 = w_5 = w_6 = 1$, $b_1 = b_2 = e$

$x_1 = 1.2$, T_1 gpa $= 3.0$

$x_2 = 0.8$, T_2 gpa $= 2.5$

$x_3 = 2.0$, T_3 gpa $= 4.0$

Forward Pass

$$n_1 = \sum_{i=1}^{3} w_i x_i + b_1$$

$$= 1.(1.2) + 1.(0.8) + 1.(2.0) + 2$$

$$= 6$$

$$\text{Out } n_1 = \text{Max}(0, 6) = 6$$

Now let's calculate y_1,

$$y_1 = n_1 . w_4 + b_2$$

$$= 6.1 + 2$$

$$= 8$$

$$y_2 = n_1 . w_5 + b_2$$

$$= 6.1 + 2$$

$$= 8$$

$$y_3 = n_1 . w_6 + b_2$$

$$= 6.1 + 2$$

$$= 8$$

$$E_{total} = \sum_{u} \frac{1}{2}(t_u - y_u)^2$$

$$= \frac{1}{2}\left((\gamma-3)^2 + (\gamma-2.5)^2 + (\gamma-4)^2\right)$$

$$= \frac{1}{2}\left(25 + 30 - 25 + 6\right)$$

$$= 35.625$$

Error at $w_4 = \frac{\partial E_{total}}{\partial w_4}$

~~$\frac{\partial E_{total}}{\partial w_4}$~~

$E_{total} = \frac{1}{2}(t_1 - out_{y_1})^2 + \frac{1}{2}\sum(T_i - out_{y_i})^2$

Error at $w_4 = \frac{\partial E_{total}}{\partial out_{y_1}} \times \frac{\partial out_{y_1}}{\partial y_1} \times \frac{\partial y_1}{\partial w_4}$

$\frac{\partial E_{total}}{\partial out_{y_i}} = -(t_i - out_{y_i})$

$\frac{\partial E_{total}}{\partial out_{y_i}} = out_{y_i} - T_i$

$\frac{\partial E_{total}}{\partial out_{y_1}} = \gamma - 3 = 5$

$$\frac{\partial \text{out}_{y_1}}{y_2} = 1$$

$$\frac{\partial y_1}{\partial w_4} = \text{out}_{H_1} = 6 \qquad \frac{\partial y_2}{\partial w_5} = 6, \frac{\partial y_2}{\partial w_6} = 5$$

$$\frac{\partial E_{total}}{\partial w_4} = (5) \times 1 \times 6$$

$$= 30$$

$$\frac{\partial E_{total}}{\partial w_5} = (8 - 2.5) \times 1 \times 6$$

$$= 33$$

$$\frac{\partial E_{total}}{\partial w_6} = (8 - 4) \times 1 \times 6$$

$$= 24$$

$$w_4 = w_4 - \alpha \frac{\partial E_{total}}{\partial w_4}$$

$$= 1 - 0.01 \times 30$$

$$= 0.7$$

$$w_5 = 1 - 0.01 \times 33 = 0.67$$

$w_6 = 1 - 0.01 \times 24$

$= 0.76$

$$\frac{\partial E_{total}}{\partial w_1} = \frac{\partial E_{total}}{\partial out\, n_1} \times \frac{\partial out\, n_1}{\partial n_1} \times \frac{\partial n_1}{\partial w_1}$$

$$\frac{\partial E_{total}}{\partial b_1} = \frac{\partial E_{total}}{\partial out\, y_1} \times \frac{\partial out\, y_1}{\partial y_1} \times \frac{\partial y_1}{\partial b_2}$$

$$+ \frac{\partial E_{total}}{\partial out\, y_2} \times \frac{\partial out\, y_2}{\partial y_2} \times \frac{\partial y_2}{\partial b_2}$$

$$+ \frac{\partial E_{total}}{\partial out\, y_3} \times \frac{\partial out\, y_3}{\partial y_3} \times \frac{\partial y_3}{\partial b_3}$$

$= 5 \times 1 \times 1$

$+ 5.5 \times 1 \times 1$

$+ 4 \times 1 \times 1$

$= 14.5$

$b_2 = b_2 - \alpha b_{i_2} = 1 - 0.145 = 0.855$

B)

(b) Here is the expression for classification rule in this case.

$$y = \begin{cases} 1, & \text{if } \sum_{i=1}^{n} \alpha_i y_i e^{\frac{-|x_i - x_j|^2}{2\sigma^2}} + b \geq 0 \\ -1, & \text{otherwise} \end{cases}$$

Derivation:

Let's say we have $\{x_i, y_i\}$, $i \in [0, n]$ as our samples.

After solving optimization problem, we get,

$$w = \sum_{i=1}^{n} \alpha_i y_i \phi(x_i)$$

When we do prediction,

$$y = \begin{cases} 1, & \text{if } w^T \phi(x_i) + b \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

The gaussian kernel is like as given below.

$$K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$$

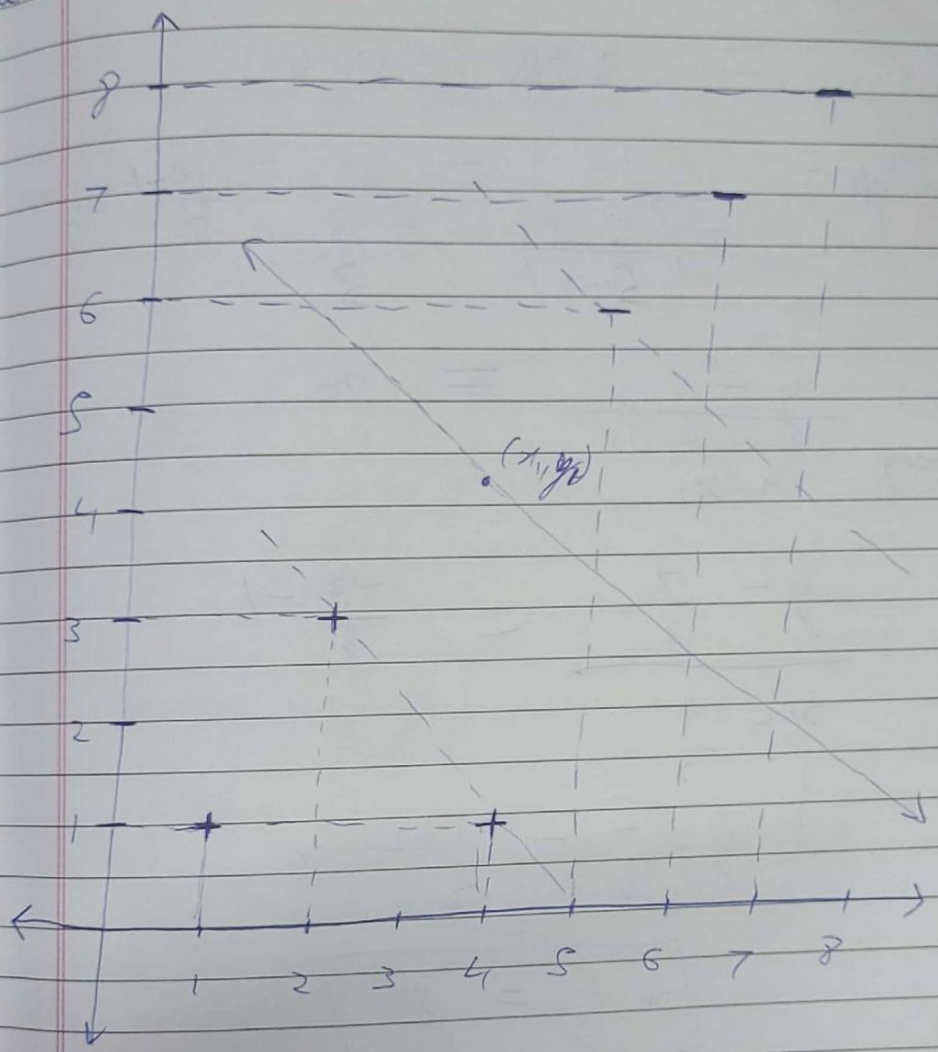
$$= K(x_i, x_j)$$

$$= e^{\frac{-\|x_i - x_j\|^2}{2\sigma^2}}$$

$$\therefore y = \begin{cases} 1, & \text{if } \sum \alpha_i y_i e^{\frac{-\|x_i - x_j\|^2}{2\sigma^2}} + b \geq 0 \\ -1, & \text{otherwise} \end{cases}$$

C)

(a)



Yes the data is linearly seperable as we can clearly see from the plots

(b) We can determine that the plot that the decision boundary will be the perpendicular bisector of (2, 3) and (5, 6) as it will have ~~maximum~~ max. distance from them.

$$(x_1, y_1) = \left(\frac{2+5}{2}, \frac{3+6}{2}\right)$$

$$= (3.5, 4.5)$$

$$\text{Slope} = \frac{1}{-\left(\frac{6-3}{5-2}\right)} = \frac{1}{-\frac{3}{3}} = -1 = m$$

Decision boundary: $\frac{y - y_1}{x - x_1} = m$

$$\frac{y - \frac{7}{2}}{x - \frac{9}{2}} = -1$$

$$y - \frac{7}{2} = \frac{9}{2} - x$$

$$\boxed{x + y = 8}$$

(c) The support vectors are (4,1), (2,3) and (5,5). They are the closest datapoints on either side.

(4,1), (3,3) lies in || to the decision boundary

(d) Margin = Distance between support vector and hyperplane

= Distance between opposite support vectors

= Distance b/w (2,3) and (5,6)

$= \sqrt{(5-2)^2 + (6-3)^2}$

$= \sqrt{3^2 + 3^2}$

$= 3\sqrt{2}$

$= 3 \times 2.121 = 6.363$

(a) If we remove any one of the support vectors, the optimal margin would change and readjust to nearest distance from new support vectors

(c) If we remove either (2,3) and (4,1) the optimal margin will not change.

If we remove (5,5) optimal margin will change.