$P_1$ $P_2$ $P_3$ $P_4$

Q1. $D = \{(0,1,1), (0,2,1), (4,0,-1), (5,0,-1)\}$

Initial seed points $\Rightarrow C_1 = (0,0)$
$C_2 = (3,0)$

Distance of $P_1$ from $C_1$: $\sqrt{(0-0)^2 + (1-0)^2} = 1$ ←
from $C_2$: $\sqrt{(0-3)^2 + (1-0)^2} = \sqrt{10}$

$P_2$ from $C_1$: $\sqrt{(0-0)^2 + (2-0)^2} = 2$ ←
from $C_2$: $\sqrt{(0-3)^2 + (2-0)^2} = \sqrt{13}$

2

$P_3$ from $C_1$ : 4
$P_3$ from $C_2$ : 1 ←

$P_4$ from $C_1$ : 5
from $C_2$ : 2 ←

b) Pts. closest to $C_1$ : $\{P_1, P_2\}$
New value of $C_1$ after update : $\frac{(0,1) + (0,2)}{2}$

$= (0, 1.5)$

Pts. closest to $C_2$ : $\{P_3, P_4\}$
Updated value of $C_2$ : $\frac{(4,0) + (5,0)}{2} = (4.5, 0)$ 1

c) Purity of clustering :

Cluster 1 : # points in class $y_e = 1$ : 2
# points in class $y = -1$ : 0
# total point / size of cluster = 2

0.5

$P_{1,1} = \frac{2}{2} = 1$

$P_{1,-1} = 0$

M T W T F S S

Page No.: YOUVA

Date:

purity of cluster 1 = $\max(P_{1,+1}, P_{1,-1}) = 1$.

Cluster 2: # pts in class $y=1$ : 0

# pts in class $y=-1$ : 2

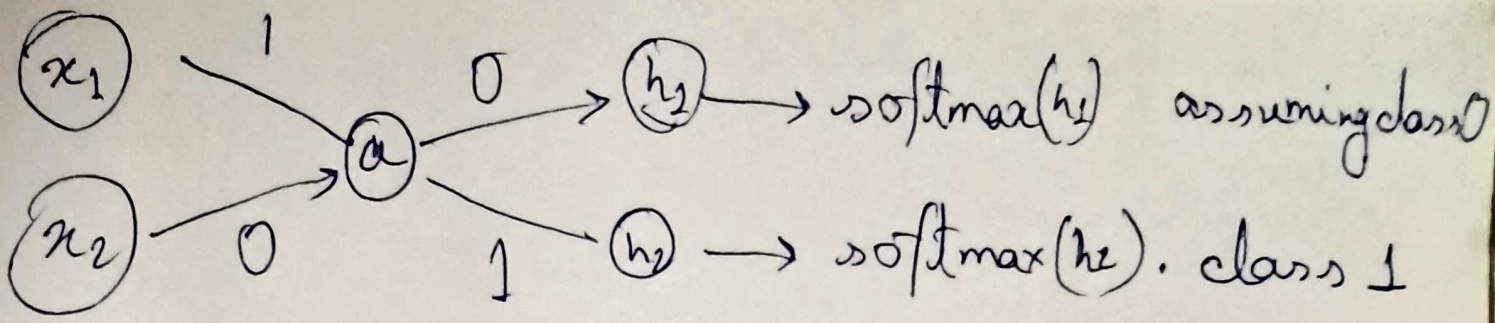
0.5

Similarly purity of cluster 2 = 1

Purity of clustering = $\frac{1}{2} \times (\text{purity}_{\text{cluster 1}} + \text{purity}_{\text{cluster 2}})$

$= \frac{1}{2} \times (1+1) = 1.$ 0.5

12.)




$h1 = (1 * x_1 + 0 * x_2) * 0 = 0$

$h_2 = (1 * x_1 + 0 * x_2) * 1 = x_1$

for data (1, 1, 1), $x_1 = 1$, $x_0 = 1$, $y = 1$

$$\text{softmax}(h_1) = \frac{e^0}{e^0 + e^{x_1}} = \frac{1}{1+e} = 0.27$$

$$\text{softmax}(h_2) = \frac{e^{x_1}}{e^0 + e^{x_1}} = \frac{e}{1+e} = 0.73$$

ii) Crossentropy for multiclass →

$$= -\sum_{i=0}^{C} y_i \log(p_i) \quad y_i \text{ is ground truth}$$

here for [0, 1] $L = -\sum_{i=0}^{1} y_i \log(p_i)$

$L = -0 * \log(0.27) - 1 * \log(0.73)$

$L = 0.13$

c.) $a = 1 * x_1 + 0 * x_2$

$= x_1$

So for batch normalization for 2 data points $(1, 1, 0), (-1, -1, 1)$ we calculate

mean $\mu = \frac{1-1}{2} = 0$ (1)

variance $= \frac{(1-0)^2 + (-1-0)^2}{2-1} = 2$ (1)

b.) MSE loss for $(1, 1, 0)$ (0.5)

$h_1 = 0, \quad h_2 = x_1 = 1$

MSE loss $= (h_1 - x_1)^2 + (h_2 - x_2)^2$

$= (0 - 1)^2 + (1 - 1)^2$

$= 1$ (1)

Sol$^n$ ① We have,

$$L(y, F(x)) = e^{.5(y - F(x))^2}$$

Differentiating w.r.t $F(x)$ we get-

$$\Rightarrow \frac{\partial L(y, F(x))}{\partial F(x)} = -\left(e^{0.5(y-F(x))^2} * 0.5 * 2(y - F(x))\right)$$

$$= -\left(e^{0.5(y-F(x))^2} \cdot (y - F(x))\right) \quad [1.5]$$

$$= -L(y, F(x)) \cdot r$$

$$\text{or } -\frac{\partial L(y, F(x))}{\partial F(x)} = L(y, F(x)) \cdot r \quad [0.25]$$

where $r = y - F(x)$ is the residue.

In gradient boosting, the updated labels after each iteration are given by the negative gradient of the loss.

So, the updated labels are $\boxed{L(y, F(x)) \cdot r}$ [0.25]

Sol$^n$ ④ For a single sample $x_i$, we have the following pdf

$$p(x_i|\theta) = e^{-\theta x_i}$$

(a) ∴ Joint likelihood of observing n iid samples with the above pdf is given by

$$L(\theta) \equiv p(x_1, x_2, \dots x_n | \theta)$$

$$= \prod_{i=1}^{n} e^{-\theta x_i} = e^{-\theta \sum_{i=1}^{n} x_i} \quad [1]$$

$$\Rightarrow \log L(\theta) = \log\left(e^{-\theta \sum_i x_i}\right) = -\theta \sum_{i=1}^{n} x_i \equiv l(\theta)$$

(b) To find MLE of $\theta$, we need to differentiate $L(\theta)$ wrt $\theta$ and set it to 0.

$$\frac{\partial L(\theta)}{\partial \theta} = e^{-\theta \sum_i x_i} \left(-\sum_{i=1}^{n} x_i\right) = 0$$

we can see that this way we are not able to solve for $\theta$. [1]

Therefore, to obtain the $\theta$, that maximizes the likelihood we can apply gradient ascent here.

$$\theta = \theta + \eta \frac{\partial L(\theta)}{\partial \theta} \quad [1]$$

Q5) To make the hidden node maximally active, we determine the input $x$ that maximises the output of hidden node.

$h = \frac{1}{1+e^{-z}}$ is a montonically increasing function, so, we'll maximise $z$.

$z = w.x$, to maximise $z$ we make the angle b/w $w$ & $x$ as 0

So, $x = cw$, i.e $x$ is scalar multiple of $w$.

To make the hidden node maximally active, we keep $\|x\| = 1$ and optimise the dot product.

one can also use Lagrangian method

$c = \frac{1}{\|w\|}$

$x = \frac{w}{\|w\|} = \frac{w}{\sqrt{w^T w}}$

2

Q6) To fit a linear regression model b/w two points, we can fit a line that minimises the error.

$$\text{error} = \sum (y - \hat{y})^2$$

$$= \sum [y - (wx+b)]^2$$

$$= (w+b)^2 + (2w+b-1)^2$$

$$= 5w^2 + 2b^2 + 6wb - 2b - 2w + 1$$

Q6) To fit a linear regression model b/w two points, we can fit a line that minimises the error.

$$\text{error} = \sum (y - \hat{y})^2$$

$$= \sum [y - (wx+b)]^2$$

$$= (w+b)^2 + (2w+b-1)^2$$

$$= 5w^2 + 2b^2 + 6wb - 2b - 2w + 1$$

$$\frac{\partial(\text{error})}{\partial w} = 0 \quad \& \quad \frac{\partial(\text{error})}{\partial b} = 0$$

$$10w + 6b - 2 = 0 \qquad 4b + 6w - 2 = 0$$

$$\boxed{w = 1, \; b = -1}$$

Use, $w = (X^T X)^{-1} X^T y$ - [we will add a dimension to X with value 1 to get b in w]

$$w = \left( \begin{bmatrix} 1 & 2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 2 & 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$X^T \quad X \quad X^T \quad y$

$$w = \left( \begin{bmatrix} 5 & 3 \\ 3 & 2 \end{bmatrix} \right)^{-1} \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \frac{1}{1} \begin{bmatrix} 2 & -3 \\ -3 & 5 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

$$w = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

So, $\boxed{w = 1, \; b = -1}$

(b) Compute w & b using D1

$\boxed{w = 1,\ b = -1}$

Compute w & b using D2

$$w = \left( \begin{bmatrix} -1 & -2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} -1 & 1 \\ -2 & 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} -1 & -2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} -1 \\ 0 \end{bmatrix}$$

$$= \left( \begin{bmatrix} 5 & -3 \\ -3 & 2 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \frac{1}{1} \begin{bmatrix} 2 & 3 \\ 3 & 5 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

$$= \begin{bmatrix} -1 \\ -2 \end{bmatrix}$$

$\boxed{w = -1,\ b = -2}$ .5

$$E\left(\hat{f}(x)\right) = \frac{E\left(\hat{f}_{D1}(x)\right) + E\left(\hat{f}_{D2}(x)\right)}{2}$$

wavg = (1-1)/2

bacg = (-1+2)/2

E(fhat) = wavg x + bavg

bias = E(fhat) - f

$$= \frac{E(x-1) + E(-x-2)}{2}$$

$$= \frac{E(-3)}{2} = -\frac{3}{2}$$

$$\text{bias} = E\left(\hat{f}(x)\right) - f(x)$$

$$= \frac{-3}{2} - x$$